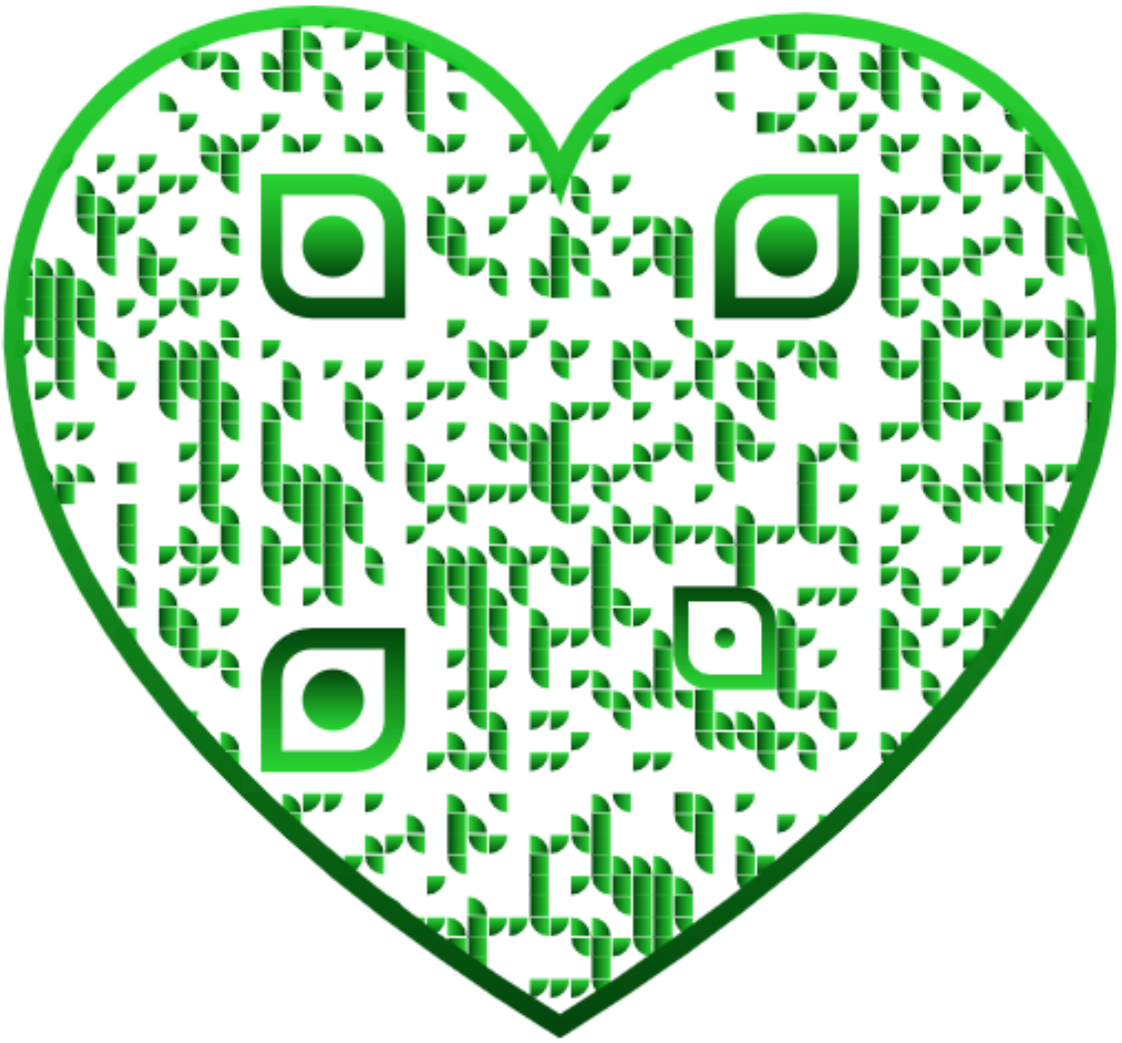
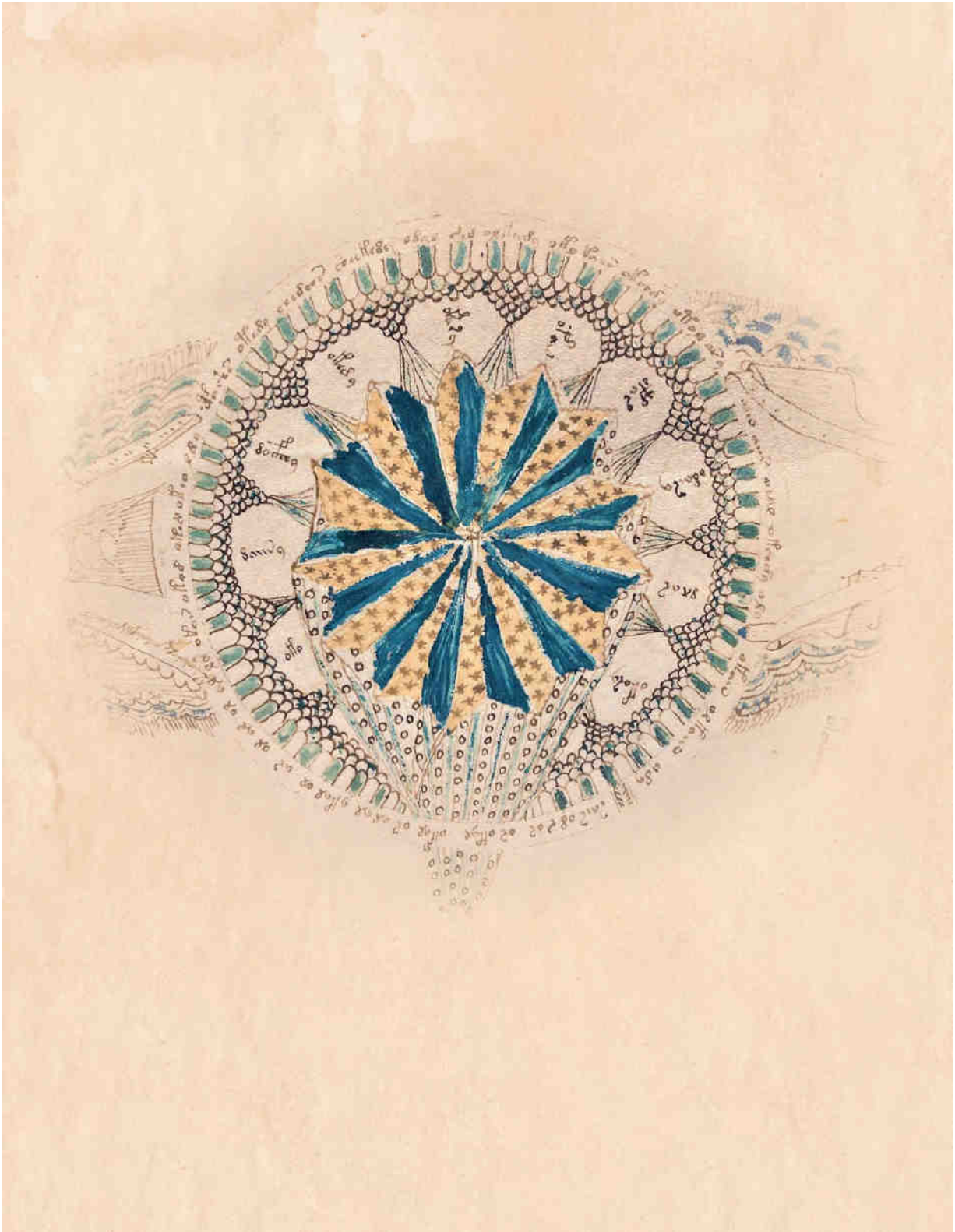
VOYNICH RECONSIDERED

The Most Mysterious Manuscript in the World

Robert H. Edwards, PhD





CONTENTS

Prologue

CHAPTER 1 **Voynich**

CHAPTER 2 **The Book**

CHAPTER 3 **The Glyphs**

CHAPTER 4 **Zipf**

CHAPTER 5 **Currier**

CHAPTER 6 **D'Imperio**

CHAPTER 7 **Markov**

CHAPTER 8 **The Gold Bug**

CHAPTER 9 **Cannabis**

CHAPTER 10 **Aries**

CHAPTER 11 **Cappelli**

CHAPTER 12 **Cantigas**

CHAPTER 13 **Illyria**

CHAPTER 14 **Voynich 2022**

Epilogue: However Improbable

PROLOGUE

This much we know, or we think we know.

Sometime in the fourteenth or fifteenth century (we cannot be more precise than that), somewhere in Europe (probably in southern Europe), a farmer or several farmers slaughtered at least fourteen or fifteen calves and sold their hides. These transactions could have occurred in one place or several, and on one occasion or on several occasions separated even by years or decades. In due course, the hides were processed into parchment or vellum,[1] and found their way, together or separately, into the hands of a merchant of writing materials. The material yielded at least 116 usable pages for a manuscript.

Five or more centuries later, scientists armed with radiocarbon dating technology would sample four pages from the parchment. They would assign the dates of death of the respective unfortunate animals to five overlapping time periods, each with a probability distribution. The five periods with 95 percent probabilities would together span the years 1308 to 1458.[2]

To put these dates into context: this period marked the Hundred Years' War between the kingdoms of France and England, but the rest of Europe was mostly at peace. Most people believed that the sun, the planets, and the stars revolved around the earth; Copernicus was, as yet, unborn. Arabia, India, and China were part of the world, as known to Europeans, but the Americas were not. In the year 1453, the Byzantine capital of Constantinople would be conquered by the Ottoman Empire. Johannes Gutenberg would perfect his technology of printing around 1440 and would produce the first European printed document by around 1450, followed in 1455 by the celebrated *Gutenberg Bible*.

Joanne Filippone Overty provides an essential backstory, which refers to medieval England but might reasonably be thought to apply to any European country of that era:

> … in England during the period 1300–1483, … manuscript production shifted from ecclesiastical institutions, such as monasteries, to more professional urban-based organisations…. Medieval book prices are not easily quantifiable, by virtue of a manuscript's nature as a unique hand-produced item. The book trade before the advent of printing was a bespoke trade, where buyers would likely contract with several different practitioners of the book arts (such as a parchment maker, scribe, illuminator, and binder), either directly or through a stationer, to purchase products that precisely met their specifications and budget.

After the parchment was made, it came into the hands of someone who wanted to make, or to commission, a profusely illustrated book. We do not know when this happened; the inks that were used cannot be carbon-dated, and the pigments were in common use for centuries.[3] We also do not know how long it took to create the manuscript: we could guess at several months, at a minimum, or possibly several years.

Eventually a manuscript was assembled. If it was written before 1440, then we can classify it as a document that predated printing in Europe, and that therefore could only have been produced by hand. It was a bespoke

project, conceived and crafted by someone for his or her own pleasure or for sale or gift to someone else.

It was packed with illustrations, at least one on nearly every page, all drawn in an ink that is now dark brown, and most of them roughly painted in blues, greens, reds, and browns (and perhaps yellows, which are now brown). The drawings seemed to be grouped into at least six sections, variously depicting plants, signs of the zodiac, alchemical objects, animals, structures, naked women (but very few men), and geometrical figures of unknown significance. Most remarkably, it contained over one hundred thousand characters or symbols that, for the most part, bore no resemblance to any script in a European language, but seemed to be arranged in strings that looked like words.

At some point the manuscript was sold to a customer or presented to a wealthy patron, we suppose the first of many.

Who that person was, we have no way of knowing. But a thirdhand report gives us some ground for thinking that, sometime in the late sixteenth century, the manuscript was in the possession of Rudolf II of the House of Habsburg, the Holy Roman Emperor and King of Bohemia, King of Hungary and Croatia, and Archduke of Austria. If that is true, the manuscript had entered the library of one of the most powerful monarchs of Europe: a man who presided over the lives of at least twenty-three million people.[4] His domain encompassed much of what is now Austria, Croatia, the Czech Republic, Hungary, Romania, Slovakia, and Slovenia, as well as parts of modern Germany, Italy, Poland, and Ukraine. If he owned the manuscript, it must have been after 1576 (when he was crowned emperor), and before 1612 when he passed away. If it was before 1583, the manuscript would most probably have been kept in his palace in Vienna; after 1583, it might have moved with his court to Prague in Bohemia.[5]

There is however no document from the court of Rudolf II, or from any contemporary historian, that confirms the emperor's possession of the manuscript.

Centuries later, under the scrutiny of ultra-violet light, a faint signature would be detected on the first page of the manuscript. It seemed to read "Jacobi a Tepenecz," followed below by several marks which could possibly be numerals. There exist today, in collections of medieval documents, at least four other books with similar signatures, also

accompanied by numerals: they are reliably attributed to Jacobus Horčický, a rich pharmacist and alchemist from Bohemia. In 1607, Horčický was appointed imperial chemist at the court of Rudolf II, and in 1608 he requested and was granted the noble title of "de Tepenec."[6] If the signature on the manuscript is his, it must have been written no earlier than 1608. We do not know how Horčický came to possess, and to sign, an item that might previously have been the property of the emperor.

Horčický died in 1622. The next known reference, or possible reference, to the manuscript occurs in a letter, probably sent in the fall of 1637. The letter, undoubtedly written in Latin, was authored, or dictated by Georgius Barschius, a legal professional at the Supreme Burgrave Court of Prague (the term "Burgrave" referring to the hereditary ruler). It was sent to Athanasius Kircher, a Jesuit priest and celebrated professor of mathematics, physics, and Oriental languages at the Collegium Romanum in Rome. The letter is now lost, but later correspondence reveals that what Barschius sent to Kircher was a handmade copy of some hitherto undecipherable characters in an antique document, accompanied by an earnest request to the famous professor to make sense of the writing. Barschius did not send the original document. We do not know for sure that this was the same manuscript that Rudolf II might have owned, and which bore the signature of Horčický; and if it was, how it had passed into the library of Barschius.

Barchius apparently did not receive a reply, because on April 27, 1639 (signing himself "Georgius Baresch") he wrote again. His letter has been preserved, and the splendid translation by Philip Neal makes it clear that this was a follow-up:

> Now since there was in my library, uselessly taking up space, a certain riddle of the Sphinx, a piece of writing in unknown characters, I thought it would not be out of place to send the puzzle to the Oedipus of Egypt [he meant Kircher] to be solved. And so, I ordered a certain old book to be transcribed in part, with the writing closely imitated (the bearer of this letter will inform you that he saw it with his own eyes). A year and a half ago I sent that writing to your Reverence.
>
> … From the pictures of herbs, of which there are a great many in the codex, and of varied images, stars and other things

> bearing the appearance of chemical symbolism, it is my guess that the whole thing is medical, …
>
> … the volume contains pictures of exotic plants which have escaped observation here in Germany.[7,8]

However, Kircher had not only received the letter and its attachments, but had already replied, not to Barschius but to Theodorus Moretus who (it may be presumed) had written the letter of 1637 on behalf of Barschius and was probably the "bearer of this letter" to whom Barschius had referred. On March 12, 1639, Kircher wrote to Moretus:

> As for the book filled with some sort of mysterious steganography which you enclosed with your letter, I have looked at it and have concluded that it requires application rather than insight in its solver…. I expect I shall try to solve it when the mood and inspiration take me.
>
> … the other sheet … was printed in the characters that people call those of D. Hieronymus; and they use the same characters here in Rome to print missals and other sacred books in the Illyrian language.[9]

It appears that Kircher saw no intrinsic difficulty in deciphering the manuscript. Intriguingly, he referred to another attachment to Moretus's letter; we do not know whether this second sheet was also dispatched by Barschius, or whether it was a project of Moretus's own initiative. A possible interpretation of the letter is that the two attachments were in the same script, namely that of Hieronymus, later to be known as St. Jerome. If that were so, the subsequent four centuries of research on the manuscript may have been misdirected.

However, there is no record that Kircher ever sent any further thoughts on the subject to Barschius or to Moretus.

At some point, it appears that Barschius enlisted the help of his friend and colleague Joannes Marcus (in Bohemian, Jan Marek) Marci, who began a correspondence with Kircher that spanned several decades. In 1640, Marci wrote to Kircher:

> The Sph*nx [he meant Kircher] will understand from the attached sheet what my friend Mr. Georg Barschius wanted to have written by me.

Here we can conjecture, though Marci did not make it clear, that Marci might have attached another partial transcription of Barschius's mysterious manuscript.

Sometime between 1646 and 1662, Barschius died.[10] In the year 1662, Marci wrote that "When he [Barschius] was dying he made me the inheritor of his collections and chemical library."[11]

Regarding the manuscript about which Barschius had written to Kircher, it is clear that Marci resumed that correspondence in his own right. On August 19, 1665, he sent Kircher a letter which has become a central part of the narrative of the manuscript. In part it reads, in this author's interpretation of the Latin:

> This book was left to me by a special friend [he was surely referring to Barschius] as a testament, but as soon as I came to possess it, dear friend Athanasius, I intended to give it to you…. The possessor of that book once in a letter requested your judgment on some part of it, which he described and transmitted to you, by which he was persuaded that the rest could be read by you … Accept, then, what was owed to you long ago, as a proof of my affection for you, and break through its bars, if any, with your usual felicity.
>
> The Lord Doctor Raphael, instructor in the Bohemian language to the (then) King Ferdinand III of Bohemia, told me that the said book belonged to the emperor Rudolph, for which he presented 600 ducats to the bearer who had brought the book, but he thought the author to be Roger Bacon the Englishman.[12]

Here is the thirdhand report that Rudolf II had once owned the manuscript. The reference to Dr. Raphael is reliably interpreted to mean Raphael Sobiehrd-Mnišovský de Sebuzín & de Horstein, a Bohemian lawyer who had been tutor to Ferdinand III[13] (who from 1627 onward was

king of Croatia and Bohemia, presumably the title to which Marci's letter referred). The "he" who thought the author had been Bacon is unclear; in this author's understanding of the Latin, it could have been either Dr. Raphael or the Emperor Rudolf II.

Again, there is no record that Kircher ever replied to Marci on the subject of the manuscript, nor that he ever seriously studied the document.

Two and half centuries would now pass, during which period the ownership and location of the manuscript are unknown to us. The story is taken up by the man after whom the document is now known: the antiquarian book trader Wilfrid Voynich.

On April 20, 1921, Voynich presented a paper to the College of Physicians of Philadelphia. Why physicians? We can surmise that they were wealthy and that they were potential customers. The title of the paper was "A Preliminary Sketch of The History Of The Roger Bacon Cipher Manuscript."[14] This title defines Voynich's narrative: that he perceived, or wished his audience to perceive, firstly that the manuscript was enciphered (and therefore, was not written in a natural language); and secondly that the author was, as Marci's letter of 1665 had suggested, the great English philosopher Roger Bacon (ca. 1219–ca. 1292).

We can imagine the audience enthralled by Voynich's breathless presentation of his discovery:

> In 1912, during one of my periodic visits to the Continent of Europe in quest of rare old books and manuscripts, I came across a most remarkable collection of precious illuminated manuscripts. For many decades these volumes had lain buried in the chests in which I found them in an ancient castle in Southern Europe … As I hope someday to be able to acquire the remaining manuscripts in the collection, I refrain from giving details about the locality of the castle.[15]
>
> … my attention was especially drawn by one volume. It was such an ugly duckling … I found that it was written entirely in cipher. Even a necessarily brief examination of the vellum upon which it was written, the calligraphy, the drawings and the pigments suggested to me as the date of its origin the latter part of the thirteenth century.

> … it must be a work of exceptional importance, and to my knowledge the existence of a manuscript of such an early date written entirely in cipher was unknown.[16]

Voynich went on to describe a letter that he had found attached to the front cover, and which he claimed, inexplicably, to have overlooked at the time of acquisition. It was none other than Marci's letter to Kircher, dated August 19, 1665 (Voynich thought the year might be either 1665 or 1666), in which Rudolf II, Raphael Mnišovský, and Roger Bacon simultaneously burst into the narrative.

Voynich's presentation, subsequently published in the *Transactions of the College of Physicians*, runs to sixteen pages and reads like a detective story, in which Voynich narrates the interplay of the undoubtedly extraordinary personalities who were at various times associated with the manuscript and whom we introduced above, as well as others whom he wove into his story:

> Who could have convinced the Emperor of its tremendous importance and claim authoritatively that it was the secret work of Roger Bacon, the alchemist of alchemists? … Shakespeare says of him that he had volumes which he prized more than his dukedom, and calls him Prospero, but in everyday life he was known as Dr. John Dee, a very romantic and mysterious personality.
>
> … in the collection of Bacon manuscripts, which unquestionably came into his possession as early as 1547, he found the cipher manuscript.[17]

In Voynich's spellbinding oration, laced with courtly and archaic phrasing and with references to distinguished persons long departed, do we hear echoes of J. B. Ward and his story of the Beale Cipher?

Much later, radiocarbon dating would dispel the possibility that Bacon could have been the author. It would also become apparent that Voynich's story of the discovery was somewhat imaginative. The manuscript had not been concealed or forgotten but had been openly offered and listed for sale in 1903, as part of a collection of the Jesuit library at the Collegium

Romanum. Voynich eventually acquired a part of this collection. The "castle," which Voynich did not wish to name, was apparently a Jesuit college in Frascati, Italy.[18]

Be that as it may: our interest is not to question Voynich's motivations nor his eccentricities (of which more in a later chapter). The manuscript itself is the star of our story. What we propose is a voyage of discovery, in which we will search for the meaning of the manuscript, if meaning there is.

In this prologue, it remains only to briefly record the subsequent history of what we may now call the Voynich manuscript, from 1921 to the present day. Voynich never sold the manuscript, and if he ever put a price on it, there is no record of how much he asked. He died in 1930, and his widow Ethel Lilian Voynich inherited his business and his assets, including the manuscript and presumably Marci's letter of 1665. She kept the manuscript until her death in 1960; in her will, she bequeathed it to her friend Anne M. Nill, who had been Wilfrid Voynich's secretary and her own lifelong companion. The circumstances of Voynich's acquisition of the manuscript then came to light; that story has been well told elsewhere, notably by René Zandbergen.

On July 12, 1961, Ms. Nill sold the manuscript for $24,500 to Hans Peter Kraus, a book trader based in New York. In 1969, the Voynich Manuscript took the final leg of its journey, when Kraus donated it to the Beinecke Rare Book and Manuscript Library at Yale University. There it resides today.[19]

With this preamble, our story begins.

CHAPTER 1

VOYNICH

Of Voynich the man, this is what we know;
or what we think we know.

Figure 1.1. Wilfrid Voynich, ca. 1885. *Courtesy of Rafal T. Prinke, public domain*

Voynich was not one to set the record straight. He wove many colorful tales of his adventures, some of them patently works of his own imagination. So, if we are to understand his life and its bearing on his eponymous manuscript, we would do well to stick close to the source documents. In this respect, we are deeply indebted to Colin MacKinnon's original research and assembly of the few official records that attest to key events in Voynich's life.

Voynich would testify that he was born Wilfrid Michael Habdank Woynicz on October 31, 1864, as a natural-born subject of the Russian Empire.[1] His birthplace was Telschi, a small town on the shore of Lake Mastis in the Russian imperial province of Kovno. The year before his

birth, Telschi had been one of the main centers of the January uprising, which pitted Polish-Lithuanian separatists against the Russian government. Today, the town is known as Telšiai and is part of Lithuania.

Voynich's parents were Leonard Habdank Woynicz and Emilia Wejciulenicz [spelling uncertain], whom he would later characterize as, like himself, Poles by ethnicity and subjects of the Russian Empire.[2]

The Russian government would later classify him as a member of the nobility, *iz dvoryan*.[3]

Of Voynich's childhood and adolescence, we know next to nothing. Since he was an ethnic Pole, we can infer that he spoke Polish at home; and as a subject of the Russian Empire, he would have spoken Russian at school. However, his later tangles with the Russian authorities left records that on November 27, 1884, aged just twenty, he had earned the degree of pharmaceutical assistant with honors, at the University of Moscow.[4] To earn a place at this renowned university, he must have had exceptional talent or connections.

He attracted official attention on October 10, 1885, when he was arrested in Kovno on the charge of membership of International Social Revolutionary Party "Proletariat," a socialist party which advocated proletarian internationalism and consequently, in the Tsarist era, was illegal. The historian Vladimir Pavlovich Skorokhodov would later write in the Russian state archives as follows:

> The files [of the Irkutsk region] preserved the track record of M. Voynich's pharmaceutical activities, compiled by the Warsaw Medical Council in 1886, from which it is clear that Voynich often moved from one city to another (Kovno, Minsk, Vilno, Warsaw). It was in these large cities that the Proletariat party concentrated its activities, on whose instructions Voynich carried out numerous assignments…. even before his arrest, Voynich was under police supervision since there are two notes from the police authorities in the record "Certificate of the local police for the time during which he did not study in pharmacies."[5]

Voynich was detained in Warsaw without trial. On April 30, 1887, after eighteen months of detention and still without trial, he was sentenced by an order of the Tsar to five years' exile in Siberia. It appears that his first residence in exile would be the city of Irkutsk in eastern Siberia, about 6,500 kilometers by road from Warsaw.

Voynich would not forget Irkutsk. The city has a coat of arms depicting a Siberian tiger, holding in its jaws an unfortunate small animal, possibly a sable. Much later, Voynich would adapt this coat of arms into the stationery of his antiquarian book business.[6]

The police records of Irkutsk, as compiled by Skorokhodov, inform us that on December 7, 1887, "having received subsistence, as a prisoner from the nobility, of 15 kopecks a day – 45 kopecks for three days," Voynich was escorted from Irkutsk to the village of Tunká near the border with Mongolia. During his time in Tunká, he evidently petitioned the governor-general of Irkutsk to return to the big city. On June 21, 1889, he received that permission. The governor-general wrote (from his mode of address, evidently to a higher authority holding government Grade VII or VIII):

> Vice-director of the Police Department Saburov, by telegram from June 21st, tells me that Mikhail Voynich, who is under police supervision in the Irkutsk province, is allowed to enter the service in Irkutsk at the Mikheevskaya hospital, according to this petition. Let it be known; I propose to Your High Well Born to declare the permission of the Police Department to Voynich.[7]

Voynich stayed in Irkutsk for nearly a year, during which time he apparently made several further petitions to the governor-general. On the grounds of ill health, he asked to be transferred to Balagansk, 285 kilometers northwest of Irkutsk, and correspondingly closer to the Arctic Circle, but supposedly with a better climate. On June 12, 1890, he was again authorized to move.

Voynich never showed up in Balagansk. On July 6, 1890, the Balagansk police wrote to Irkutsk:

> Voynich “nyet” [“no”]; please do not hesitate to make a legal search. Policeman Malkov.

The Irkutsk police searched for him for weeks; the manhunt was personally authorized by the governor-general and classified as “secret, highest importance”; but Voynich was gone.[8] We imagine that he took trains through Russia, Germany and France or Belgium, and a boat or ferry to England (he himself had a different story, undoubtedly fictional). After several months of travel, he arrived in London on October 5, 1890.[9] Skorokhodov specifies that it was “an October evening” (how would he know that?). Voynich had been imprisoned or exiled for almost five years. He was nearly twenty-six years old.

In the Russian state archives, Skorokhodov adds[10] (but one wonders: how did they know?) that on his arrival in London, Voynich went straight to the apartment of Sergey Mikhaylovich Stepnyak-Kravchinsky, the Russian revolutionary who in 1878 had assassinated General Nikolai Mezentsov, the head of the Russian secret police.

Skorokhodov finally records that it was in Stepnyak-Kravchinsky’s apartment that Voynich met Ethel Lilian Boole, a British novelist, whom he would later marry.

Of Voynich’s life in England, we have his own recollections with their inevitable inconsistences, supplemented by secondhand stories. There are a few events that we know with certainty: including his enumeration in the 1901 census; his marriage to Ethel Lilian Boole in 1902; and his naturalisation as a British citizen in 1904.

In the census of England and Wales in 1901, over ten years after his arrival in London, Voynich was recorded as living with Ethel Lilian Boole at Flat 20, Great Russell Mansions, 60 Great Russell Street, London. Then as now, it was a posh address, directly opposite the British Museum. They were rich enough to have a resident servant at the same address. Voynich gave his occupation as “Bookseller & Bibliographer” and his birthplace as “Russian Poland.”[11]

Figure 1.2. Wilfrid Voynich, ca. 1900. *Public domain*

On September 6, 1902, Voynich married Ethel at the Register Office of St. Giles in London. He gave his age as thirty-eight (he would be thirty-eight in October 1902), and his occupation as "Analytical Chemist" (why not bookseller?). His address was still Great Russell Mansions. He had elaborated and hyphenated his father's name: Leonard Antoni Fulgenciusz [spelling uncertain] Habdank-Wojnicz (where did these Italianate names come from?).[12]

On March 16, 1904, Voynich submitted his application for naturalization as a British citizen. He gave his name, without the hyphen, as "Wilfred Michael Habdank Woynicz, commonly known as Wilfred Michael Voynich." His cover letter was headed with his business address: 1, Soho Square, London. He gave his business activity as "antiquarian bookseller." Here for the first time, we learn where he claimed to have been living:

- December 1896 to September 1899, at 92 Edith Grove, Chelsea, London
- September 1899 to September 1900, at "Boarding houses in London and traveling in Italy"
- September 1900 to March 16, 1904, at 20, Great Russell Mansions, London.

He thereby claimed, as required by law, that he had had at least five years of residence in the United Kingdom during the eight-year period prior to his application. He was supported by five distinguished British referees: the current Keeper of Printed Books at the British Museum, the former (by then retired) Keeper of Printed Books, a prison commissioner, a physician, and a former public works engineer.

Voynich received his certificate of British naturalisation on April 25, 1904.[13]

Of his journey to Europe in 1912, during which, as he would later claim, he rediscovered what he would call the "Roger Bacon Cipher Manuscript," there is no official record.

In July 1914, World War I commenced in Europe. Within six months of the outbreak of war, in other words in late 1914 or early 1915, as far as we can tell, Voynich made the first of many trips to the United States. Our story of his visits to America has two strands: firstly, Voynich's efforts to enlist American interest in the "Roger Bacon Cipher," possibly with a view to decipherment or sale of the manuscript; and secondly the extraordinary official and semi-official investigation that targeted him as a suspected agent of Germany.

Figure 1.3. Wilfrid Voynich at his bookshop in Soho Square, ca. 1900–1910? *Campbell & Gray*

The investigation was triggered by at least two informants. On June 25, 1917, Voynich wrote from his office at Aeolian Hall, 33 West 42nd Street, New York, to W. S. Booth, a writer based in Boston. Voynich's letter appeared to be couched in entirely innocent terms:

> Do you remember that you took great interest in my Bacon cypher MS. I wonder what the result was of your attempts to decipher it. Now your War Department is working on the subject, and I hope they will be able to transliterate it, with the help of their expert readers.[14]

Booth promptly wrote to the Department of Justice in terms that were at once exonerating and insinuating:

> I have no reason to suppose that Mr. W. M. Voynich is not solely occupied with his bookselling, but the enclosed letter from him to me may be worth a moment's consideration in view of your reported difficulties with spies.
>
> … he is a well-known dealer; … he told me he brought "practically" his whole fortune over here from London in the shape of books and manuscripts; … he appeared here, as I remember, some six months after the war had opened in Europe.
>
> … there is a possibility that the supposedly Roger Bacon M.S. is being used as a means of communications with some of the [government] departments.[15]

As far as official records show, the Department of Justice took no action on Booth's letter. But later that year, a more serious allegation surfaced from within the Department itself: specifically in the Bureau of Investigation (the BoI, forerunner of the FBI). On November 22, 1917, Charles Daniel Frey, whom we suppose to have been an agent at the Chicago office of the BoI, wrote as follows:

> During the past year we have had in Chicago a man by the name of Wilfred M. Voynich … He is ostensibly a dealer in fine old manuscripts…. When he was here last year, he was entertained by the Librarian [the reference was to Walter Lichtenstein, the German-born Head Librarian at Northwestern University] at his house in Evanston … During the dinner he made the statement that he had in his possession the cipher of the American War Department.[16]

The Bureau of Investigation opened a file on Wilfrid Voynich. On December 18, 1917, agent A. W. Willett from the New York City office interviewed Voynich, and was both impressed and reassured:

> I called on Mr. de Voynich [sic] at his office in the Aeolian Hall Building…. He is about sixty-four years old. [Voynich was in fact aged fifty-three at the time.] … He denies absolutely ever having made the statement that he possesses the American War Department cipher, but does say, however, that he has worked hard thinking that this manuscript would be of some service to the United States Government for cipher purposes…. He opened all of his papers to me very gladly, and his story so far as it went was plausible.[17]

Further support for Voynich came from Colonel R. H. Van Deman, chief of the Military Intelligence Section at the War Department. The BoI had evidently lent the Voynich file to Van Deman. Writing on January 7, 1918, to A. Bruce Bielaski, chief of the BoI, Van Deman said:

> As to Voynich's clpher manuscript, I am bringing copies of it to the War College, and Lt. Yardley and I are going to work on it. I do not think Mr. Voynich ever asserted that it was in the possession of the War Department. If he thought it could be of any service to the Department, it was only because it appears to be an excellent cipher, inasmuch as it has for several years defied the efforts of experts to decipher it, and he thought it might be, as it undoubtedly would [be], of value to the War

> Department to possess so valuable a system, but obviously it cannot be placed at the disposal of the Government until we succeed in ascertaining what the system is.
>
> If Mr. Voynich said that he was acquainted with the army cipher, he said nothing more than any non-military man could have said who had read the published statements to the effect that the army was using the cipher disc. This has been a matter of common knowledge for many years.

Colonel Van Deman was wrong in one respect: Voynich had indeed said (in his letter to Booth, which was on file with the BoI), that the manuscript was in the possession of the War Department. Furthermore, Van Deman acknowledged that he had copies of it.

The documents in the BoI file are now declassified and are held by the National Archives and Records Administration. They span the period from July 1917 (when Booth submitted his insinuations against Voynich) to October 1920 (by which time the case was effectively closed). No action was ever taken against Voynich with regard to the allegations against him.

The episodes documented in the BoI file show how Voynich inspired strong feelings of admiration, or for those who felt otherwise, suspicion bordering on hatred and racism:

> "… distinctly a boaster, and a good salesman, and he told fascinatingly mysterious stories of his adventures in search of hidden or lost libraries" (W. S. Booth)
>
> "… quite intellectual and evidently very wealthy … a man of great intellect and of very suave manner" (Willett, BoI)
>
> "I … have the highest opinion of him as a man of honor and integrity" (A.W. Pollard)
>
> "He has put it over [the expression seems to mean deceived] most of the wealthy Chicagoans … he has not an honest hair on his head." (Charles D. Frey, BoI)
>
> "… [I had] known him for a long while and did not want to do business with him…. {I] did not trust him, and did not think him straight" (George E. Smith, book trader)

> "… [he] had a feeling of distrust for De Voynich [sic] … simply that he did not believe the man." (unnamed book trader).

Perhaps we should leave it to Stefan Juszczynski, a Polish socialist based in London, to synthesize Voynich's complex character:

> He had exuberant phantasy and took its results for reality, in which he solemnly believed. Later he became … a very practical antiquarian books dealer and made a considerable fortune, which he was always happy to share with anyone. And so in that man lived in agreement—incredible phantasy (others call it lies), truly American pragmatism and good heart.[18]

In passing, we may note that during Voynich's stays in America, some of his acquaintances referred to him by the noble-sounding name "De Voynich." Whether they or he invented this appellation, we do not know. Certainly a few Polish noble families had used "De" or "de" as part of their surnames; but to the best of our knowledge, no official document ever attached this epithet to Wilfrid Voynich.

Voynich, meanwhile, continued his quest for the decipherment of the manuscript. Booth evidently did not reply to his enquiry. Voynich however had an ally in the War College: Captain John Manly of the Cipher Bureau, Military Intelligence Section. In December 1917, Manly was quick to defend Voynich against his accusers in the BoI, and in doing so he disclosed that he himself had been working on the decipherment:

> There is no ground for suspicion in regard to Mr. Voynich's cipher manuscript. It was written in England over five hundred years ago [he meant by Roger Bacon; in this respect he was wrong] … Mr. Voynich obtained the manuscript some ten years ago when he purchased the library of one of the mediatized royal families of Europe for the lump sum—as is reported—of five million dollars. The manuscript itself he values at about $100,000. We are very anxious to have it read and have given photostat copies of several pages of it to various cipher experts

> in England, France, and America. I received seven pages from him last winter and have worked at them a good deal, without success as yet.[19]

Manly had undoubtedly bought into one of Voynich's embroideries: the collection came from the Jesuits, not from any royal family, and cannot conceivably have commanded a price of $5 million. The manuscript was eventually sold (in 1969, by Anne Nill to Hans Peter Kraus) for $25,400. But we know now that as early as the winter of 1916/17, Voynich had shared partial copies of the manuscript with the War Department, and that his statement to that effect, in his letter to Booth, was true.

Voynich would continue his focus on the American market for several years. We recall from our prologue his riveting presentation in Philadelphia in April 1921; though it seems not to have yielded a prospective buyer nor any serious attempt at decipherment of the manuscript. Voynich would make at least five more visits to America: the records of Ellis Island immigration show that he arrived in New York from England on October 22, 1922; on November 21, 1923; and on November 14, 1924; and that he returned from Boston to Liverpool on August 17, 1925, and again on August 17, 1926 (the corresponding outbound voyages were apparently unrecorded).[20]

On March 19, 1930, Voynich died in New York. The cause of death was lung cancer. He was sixty-five years old. He had not succeeded in selling his "Roger Bacon Cipher" manuscript (if indeed that was ever his intention). Nor had he made any progress toward translating or deciphering it (which he surely desired). Others would take up that challenge.

CHAPTER 2

THE BOOK

We have briefly narrated the history of the manuscript, to the extent that anything is reliably known of it; and we have introduced Wilfrid Voynich, who unwittingly gave his name to it. We need now to characterize the book itself.

On the internet and elsewhere, descriptions of the book abound: some of them in the most minute, one is tempted to say excruciating, detail. Our objective in this story is to search for meaning in the book, or otherwise to assure ourselves that there is none. Therefore, we shall confine our description of this mysterious object to the minimum that may keep us on the path to our goal.

The object itself is unassuming; Voynich described it as “an ugly duckling.” The antiquarian book dealer Hans Peter Kraus, who owned the book from 1961 until 1969, gave it a catalog entry which reads in part as follows:

> (BACON, ROGER?) Cipher manuscript on vellum. Text written in a secret script, apparently based on Roman minuscule characters, irregularly disposed on the pages. 102 leaves (of 116; lacks 14 leaves), including 7 double folio folding leaves; 3 triple folio folding leaves; and one quadruple folio folding leaf. With added signature marks (of the XVth or XVIth century), and foliation (of the XVIth or XVIIth century) … With about 400 drawings of botanical subjects, including many of full-page size; 33 drawings of astrological or astronomical subjects, plus about 350 single star-figures; and 42 (biological?) drawings, most of which include human figures. The drawings colored in several shades of green, brown, light yellow, blue, and dark red. Large 8vo (ca. 230 × ca. 160 mm.) Old limp vellum covers (now detached).[1]

The Beinecke Rare Book and Manuscripts Library, where it now resides and where it is known prosaically by the call number Beinecke MS 408, gives it this description:

- Published / created: ca. 1401–1599?
- Publication place: Central Europe [?]
- Abstract: scientific or magical text in an unidentified language, in cipher, apparently based on Roman minuscule characters.
- Description: Parchment. ff. [folios] 102 (contemporary foliation, Arabic numerals; not every leaf foliated) + I [index?] (paper), including 5 doublefolio, 3 triple-folio, 1 quadruple-folio and 1 sextuple-folio folding leaves. 225 × 160 mm.

In other words, it is a little larger than a modern paperback; it would almost fit in your pocket. It has 102 folios according to the Beinecke Library. They have been numbered at some unknown date from 1 to 116, each in the top right-hand corner, in Arabic numerals much like those we use today; but fourteen folios are missing. Ten of the remaining folios are between two and six times larger than normal and are folded so that they fit neatly within the binding.

In suggesting Central Europe as the origin of the manuscript, the Beinecke Library was probably following the narrative of Wilfrid Voynich and his proposal of the manuscript's sometime residence at the palace of Emperor Rudolf II and subsequently in the library of the courtier Jacobus de Tepenecz (therefore, in either Vienna or Prague). There is, however, no definitive documentation of the emperor's possession of the book.

What is less clear is why the Beinecke should unequivocally state, firstly that the text is written in a language; and secondly that it is in cipher (the catalog entry for the book actually gives the main title as "Cipher Manuscript" and the alternative title as "Voynich Manuscript").

Figure 2.1. The Voynich Manuscript, during one of its rare appearances in public, on April 21, 2017. *Beinecke Rare Book and Manuscript Library*

Firstly, no one has proven that the text is written in any known human language (if they had, the book would no longer be a mystery). Secondly, no one has proven that the text is enciphered. Voynich himself made this proposition, purely, as far as we can tell, on the basis of the bizarre glyphs; many analysts have accepted it. We are no more entitled to make that assumption than if, with no prior knowledge of, say, Cyrillic script or Chinese characters, we saw them for the first time.

Finally, the "Roman minuscule characters" (where evidently the Beinecke Library copied from Kraus's catalog entry) do not refer to the glyphs themselves. They refer to a theory of the American professor

William Romaine Newbold who thought he saw “irregularities of the edges of the letters [glyphs] as seen under a magnifying glass, which he converted to letters.”[2] Most Voynich researchers believe that Captain John Manly disproved Newbold’s theory in 1931. It is remarkable that the Beinecke Library appears to espouse it. Perhaps this reference was inserted by a junior curator, and no one checked it.

THE PARCHMENT

While the Beinecke Library surely knows that the parchment dates from between 1308 and 1458, it allows for the creation of the manuscript to have been up to a century and a half later than the latter date. In other words, the curators of the Beinecke can accept the idea that as late as 1599, someone, or several persons, acquired a stock of parchment from the fourteenth or fifteenth century, and caused the text and illustrations to be drawn on it. Why the end of the sixteenth century should be the cutoff date is not clear; to this author’s knowledge, no one has definitively excluded a later date, even as late as 1912. (We are not suggesting that Wilfrid Voynich created the book; no one has proved that he did, or that he did not.) The inks and pigments are inorganic, and as we observed in our prologue, they cannot be carbon-dated.

Here is what we know about the age of the parchment.

To the best of our knowledge, the book was subjected to a scientific dating process on only one occasion. The Yale University Library, in January 2009, asked the Accelerator Mass Spectrometry Laboratory at the University of Arizona to conduct that analysis. To that end, Yale provided the Laboratory with four samples, from folios #008, #026, #047, and #068, each sample weighing between 5 and 11 milligrams.[3]

The AMS Laboratory, using the well-established technology of radiocarbon dating, made five estimates of the age of the samples: one each for folios #008, #026, and #047, and two for folio #068 (with and without solvent treatment[4]). In each case the estimates had a probability distribution with two peaks; in most cases the later peak had a much higher probability than the earlier.

In February 2009, the laboratory reported back to Yale University Library. The estimated dates of the four samples were as follows:

Folio #008 (with solvent extraction) . . . . 1394 to 1458 (92.2 percent)

Folio #026 (with solvent extraction) . . . . 1391 to 1447 (83.4 percent)

Folio #047 (with solvent extraction) . . . . 1392 to 1449 (88.2 percent)

Folio #068 (with solvent extraction) . . . . 1308 to 1362 (42.5 percent)

. . . . 1385 to 1436 (52.9 percent)

Folio #068 (without solvent extraction) . . 1396 to 1457 (93.8 percent)[5]

The probability distributions of the five tests are summarized in Figure 2.2.

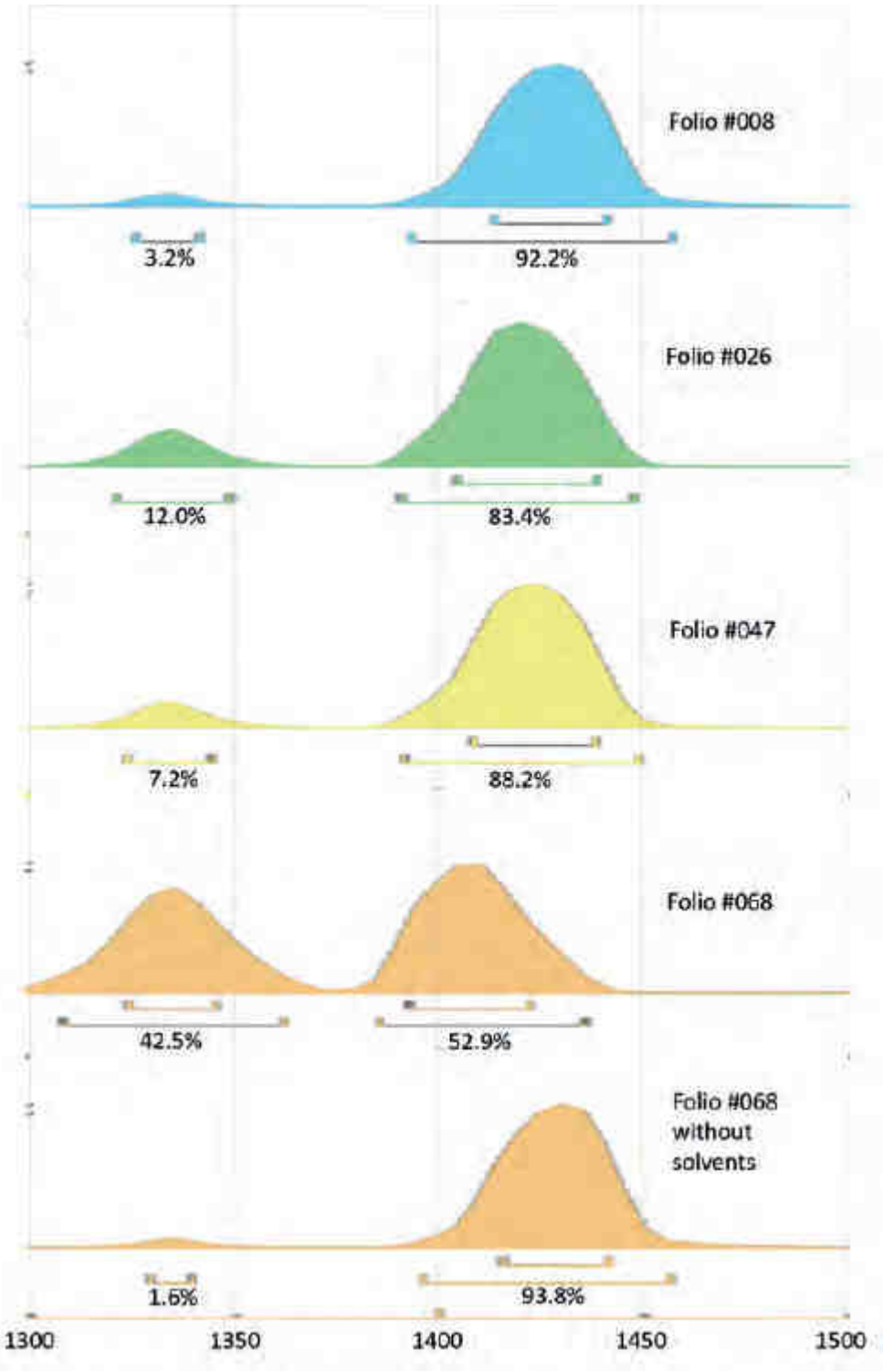


Figure 2.2. Voynich manuscript, folios 008, 026, 047, 068: probability distributions of the dates of production of the parchment, based on radiocarbon analysis, February 2009. Source: adapted by author from Hodgins, Gregory (2009), pp. 5–7. The shorter horizontal bars represent probabilities about two-thirds those of the longer bars.

Broadly speaking, these results seem to tell us that the parchment for these four folios was most probably produced sometime in the first half of the fifteenth century. The only anomaly is folio f068, which on one of its tests (the one with solvent treatment) yielded a probability above 50 percent of a date in the fourteenth century. Since parchment lasts for centuries, we suppose that a fifteenth-century merchant could legitimately sell material that was up to a century old.

Although the dates of the four samples are quite consistent, we cannot exclude the possibility that other folios of the manuscript, had they been tested, might have yielded earlier or later dates than these samples.

THE PAGES

Let us then iterate the physical properties of the object, insofar as they are relevant to our search.

The 102 folios may be thought of as equivalent to 227 pages, each page being either the front (recto) side of a normal-sized folio, the back (verso) of the same, or a booksized part (recto or verso) of a foldout folio. In Voynich research, each page is conventionally given its own code. For example, the folio f072,[6] which folds out into three parts, is given the page codes f072r1, f072r2, and f072r3 for the recto, and f072v1, f072v2, and f072v3 for the verso. On the basis of this coding, we may now describe the book in a quantitative sense.

Of the 227 pages, 195 contain substantive illustrations as well as characters, and thirty-two essentially contain only characters. The illustrations have been drawn in ink (it looks like the same ink as that of the characters), and most of them have then been colored, sometimes neatly and sometimes crudely. The pages which have only characters reveal nothing of their meaning (although there have been plenty of proposals). However, the illustrations are mostly comprehensible to the modern eye (or at least permit an interpretation); in some cases, they have annotations in what looks like French. They provide a basis for dividing the book into themes or sections. There are varying views on what constitutes a theme; we are happy to

follow the definitions on the splendid website of René Zandbergen, as follows.

There is an "**herbal**" section, consisting of 129 pages and thereby comprising more than half of the book. Of these, 110 pages are contiguous, and the rest are clumped elsewhere in the book. Each page contains a drawing of a plant or plants, including the stems, leaves, flowers or berries, and the root system (and in one case, an unidentified small animal apparently nibbling on the plant). There is no specific indication that any of the plants is an edible or medicinal herb. If there is more than one plant on the same page, they usually appear to be of the same species. The illustrations all use a palette of four or five colors (green, red, blue, a dark brown and a light brown, which might once have been yellow); the browns may be different dilutions of the same pigment.

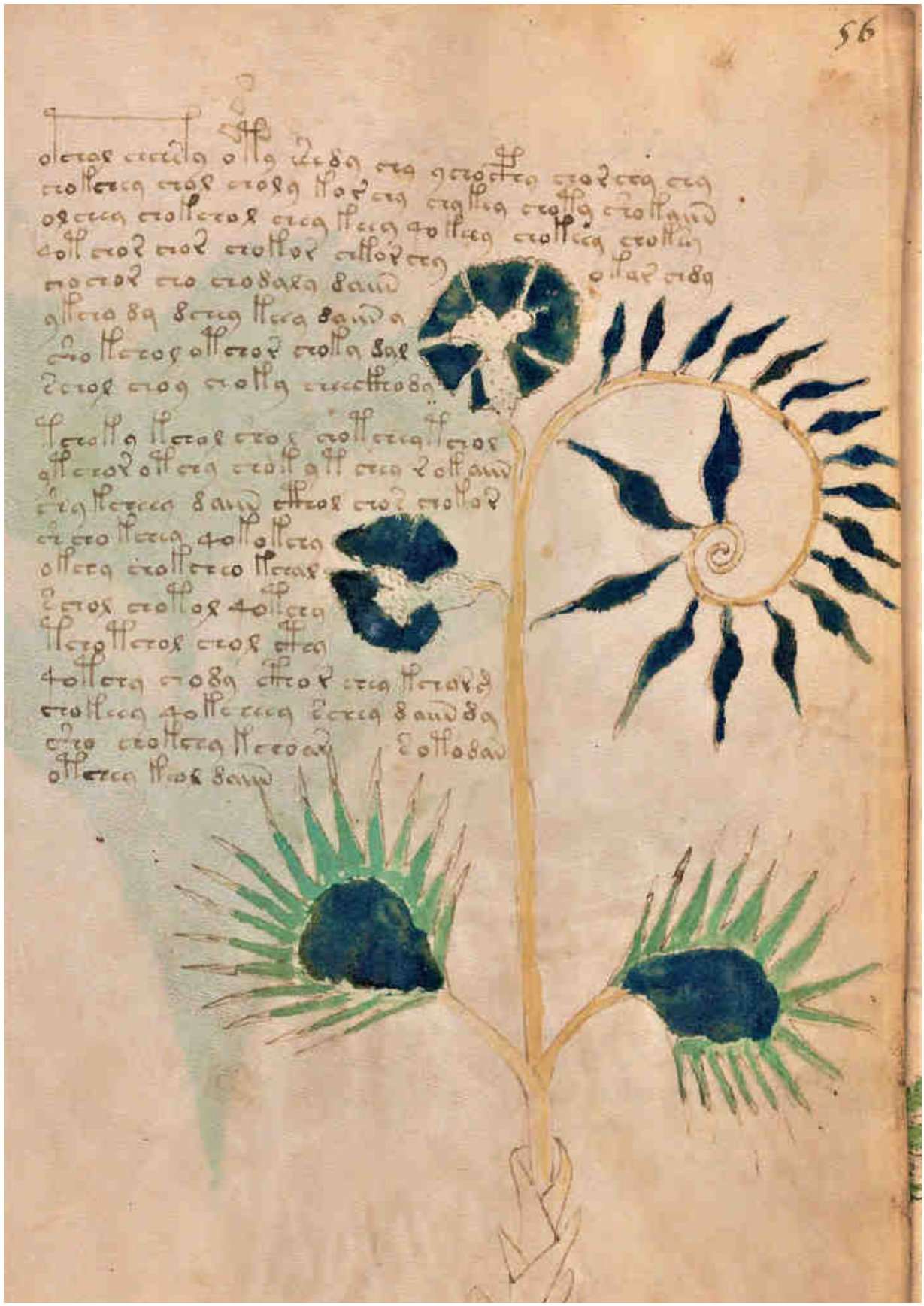


Figure 2.3. Voynich manuscript, page f056r, one of the "herbal" pages, depicting an unknown plant. The color palette includes green, light brown and blue. The drawing on the reverse of the folio (page 056v) shows faintly through the parchment. Dr. Edith Sherwood cited Denis Barthel as having identified this plant as round leaf sundew (*Drosera rotundifolia*). *Beinecke Rare Book and Manuscript Library*

It has to be said that the illustrations of plants are not naturalistic in the style of the Victorians or Edwardians. Rather, they are stylized, fanciful, one might say unearthly. There is an alien quality to them. One plant has human faces attached to its roots; another is rooted in the headless body of a cat. To this author's inexpert eye, not one of them has a familiar look. There have been many attempts to correlate them with modern species; some of them partial and tentative, and some of them (notably that of Dr. Edith Sherwood) comprehensive and with the ring of authority. Nevertheless, to this author's knowledge, there is no consensus on the identification of a single plant. This is all the more remarkable considering that, if we knew with confidence the genus or species of one plant, we would surely expect to find the name of that plant, in some language, among the characters on that page.

The plant illustrations deserve their own chapter in this book, and it will follow.

Most researchers have identified one or more sections with an extraterrestrial theme. René Zandbergen sees three sections, which he calls **astronomical**, **cosmological**, and **astrological**. They are short sections; this author is tempted to group them together into a "**cosmic**" theme, occupying thirty-one pages.

The illustrations on the "astronomical" and "cosmological" pages seem to depict the sun and the moon (both bearing human faces), circles made up of stars, and possibly constellations. There is nothing that could be identified as a planet; nor would we expect it, since it would be 1610 before Galileo Galilei would observe the extraordinary rings of Saturn. The colors in these illustrations have a more muted palette: mainly blue, often brown, occasionally red and green. The illustrations are stylized, geometric and symmetrical; they do not depict any physical relationships between the cosmic bodies, with the possible exception of a group of seven stars which some researchers have associated with the Pleiades. We may recall that in Europe in the early fifteenth century, the shapes of constellations were common knowledge; the five planets Mercury, Venus, Mars, Jupiter, and

Saturn had been known since antiquity; and the sun, moon, stars and planets were presumed to orbit the earth.

The “astrological” section has twelve pages which unmistakably depict some (but not all) of the constellations of the zodiac. Each of these pages contains an annotation, consisting of a single word in a Latin script, in a language resembling Middle French, which is surely the name of a month.

Figure 2.4. Voynich manuscript, detail from page f068r, one of the “astronomical” pages. The seven small stars in the upper left corner have been associated with the Pleiades. *Beinecke Rare Book and Manuscript Library*

Figure 2.5. Voynich manuscript, detail from page f070v2, one of the "astrological" or "zodiac" pages, depicting the sign of Pisces. The annotation "Mars" appears to be in French and means March (the month associated with Pisces). *Beinecke Rare Book and Manuscript Library*

Again, if we could correlate a single one of the "cosmic" pages (of which, the zodiac pages are the prime candidates) with a single name of a corresponding star or constellation, in any modern or medieval language, we would have a crib which might "break the bars" of the manuscript, as

Barschius put it in his letter to Kirscher. We therefore need to devote a chapter to the "cosmic" pages, and we shall do so.

Researchers generally identify a "**pharmaceutical**" section which occupies sixteen pages, in two contiguous clumps of six and ten pages respectively. The illustrations on these pages depict ornate quasi-cylindrical structures or containers, standing on their bases or legs, resembling layer cakes or Russian samovars, and accompanied by parts of plants, notably roots and leaves. The plants, being apparently incomplete, do not readily lend themselves to identification; consequently, this section does not self-evidently inspire a search for clues to the characters.

In passing, we recall from chapter 1 that Wilfrid Voynich earned the degree of assistant pharmacist in Moscow; and that on his marriage certificate, he described himself as an analytical chemist. We wonder what he made of the "pharmaceutical" pages, and whether, in his attempts to decipher the manuscript, he searched for the scientific or vernacular names of chemical elements or compounds. If he did, no record of his search remains.

Undoubtedly the most bizarre section is the "**biological**" or "**balneological**" section (the divergence of these two titles reflects the confusion, or perhaps factionalism, of Voynich researchers). It has nineteen contiguous pages. The illustrations depict naked women in what look like inter-connected pools of water, or tubs or baths of some kind, often connected by pipes or tubes of a metallic appearance. The women generally are not doing anything specific; they seem to be sitting or standing quietly in the water. There are no identifiable male figures in these drawings. The palette is predominantly green and blue; there are traces of pale brown or faded yellow; there is no red.

Figure 2.6. Voynich manuscript, detail from page f099r, one of the “pharmaceutical” pages: showing one of the “samovars.” *Beinecke Rare Book and Manuscript Library*

Those who see this section as biological adopt the narrative that the pools and tubs are metaphors for organs of the human body (perhaps the female body). The balneological epithet, on the other hand, reflects the idea that the pools and baths are real-life structures in the nature of spas or healing springs.

Figure 2.7. Voynich manuscript, detail from page f099v, one of the “pharmaceutical” pages: showing a group of objects that resemble the roots of plants. *Beinecke Rare Book and Manuscript Library*

Figure 2.8. Voynich manuscript, detail from folio f075v, one of the "biological" or "balneological" pages, depicting ten naked women in a pool or bath of green water. *Beinecke Rare Book and Manuscript Library*

There is nothing in the biological/balneological section that prompts an immediate idea of correlation with known objects or places. The pools do not clearly have the shape of human organs; the baths and pipes do not conjure an identification with any known therapeutical destination, either of medieval or of modern times.

THE LAST PAGE

Although most Voynich researchers view it as part of the "text" section, the last page of the manuscript (folio f116v) deserves a special mention. The text of this page has only four lines, which are reproduced in figure 2.9.

These four lines have engendered decades of speculation. Many researchers have conjectured that most of the text of these lines is in a European language and script, possibly Latin or German or some amalgam of the two. Several researchers have proposed that the last five strings could read "so nim geiß mich o" and that this could be German for "… take goat's milk, or …"[7] For what it is worth, Google Translate reads this phrase as "so don't kick me" or "so don't kill me."

If indeed this page contained words in any European language, it might amount to a crib: a Voynich "word" or "words" with their equivalents in a

natural language.

There is no dispute that the fourth line contains two strings of characters which occur elsewhere in the manuscript: ɑ୨oy ୡc89 or {ayoy.2c89} in the v101 transliteration (which we shall introduce in chapter 3). The string ɑ୨oy {ayoy} occurs eighteen times in the whole manuscript, but only three times as a "word." The string ୡc89 {2c89} occurs 344 times in the manuscript, of which 201 times are as a "word." Captain Prescott Currier (whom we shall meet in chapter 5) explicitly identified folio 116 as being in a Voynich language and did not qualify page f116v as containing any other language.

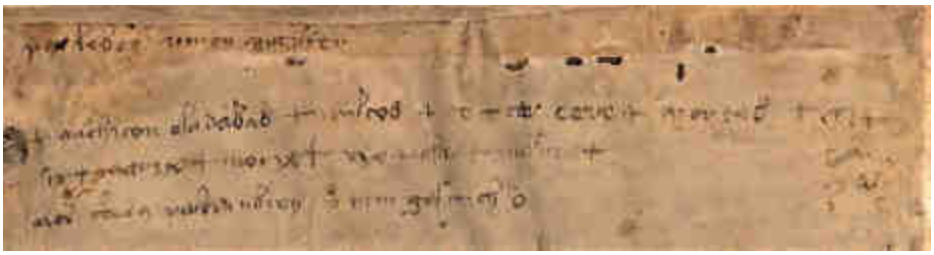


Figure 2.9. Voynich manuscript, page f116v, lines 1 to 4. (Rest of page contains no text.) *Beinecke Rare Book and Manuscript Library*

Following Occam's razor, the simplest hypothesis would be that the whole page is in Voynich characters, that some of them resemble letters in Latin script (as do many of the characters), and that others are unique to this page. We are not much inclined to search this page for a Rosetta Stone.

WHERE NEXT?

We can think of the thematic sections as falling into three groups:

- The sections which have botanical or cosmic themes admit a glimmer of hope that somewhere among their pages, there is the name of a real and identifiable plant, or a real and identifiable star or constellation, in some natural human language.

- The pharmaceutical and biological/balneological sections offer no such promise, and therefore open the door only to speculation.
- The characters as a whole admit an approach based on linguistics, powered by statistics and driven by the hope of a correlation of some kind with a known language.

PRECONCEIVED NARRATIVES

The Voynich manuscript invites, for those who are so disposed, the insertion of a preconceived narrative. In this respect, it bears comparison with the notorious proliferation of narratives relating to the man who came to be known as D. B. Cooper, and his hijacking of Northwest Airlines Flight 305 on November 24, 1971.

In the forty-five-year period, from 1971 to 2016, during which the FBI was actively investigating the case, we can guess (and often we can read in the declassified files) how often an FBI agent, we imagine tired and frustrated, had to participate in a conversation with a witness whose brother, uncle, classmate, or late husband had "done the job." The qualification of the alleged perpetrator was invariably that he (and sometimes she) had the right kind of personality to pull off a big heist. In no instance could the proposed suspect be reliably located at the scene of the crime.

Likewise, in the colorful history of the Voynich manuscript, we have the impression of reading how often someone started with such a narrative and chose the facts that suited the story. Wilfrid Voynich must bear some of the responsibility. It was he who fingered Roger Bacon as the author, and later named John Dee as the merchant or donor of the manuscript at the court of the emperor Rudolf II. Many other Voynich researchers seemed to have started with a perpetrator and to have worked backward: their suspects including Antonio Averlino, Anthony Askham, Edward Kelley, John Dee (again), Roger Bacon (again), and a youthful Leonardo da Vinci.[8] We apologize to these researchers (whose work is now often inaccessible) if in fact their reasoning was entirely driven by data and logic.

Even the proposed geographical origin and presumed languages of the Voynich manuscript have lent themselves to vastly diverse perspectives. Among the propositions have been that it was:

- written in England in medieval Latin (Wilfrid Voynich, 1921)[9]
- Latin with a complex system of encipherment (William R. Newbold and Roland G. Kent, 1928)[10]
- "(my first impression is that) it was written in the south-western corner of Europe: Spain, Portugal, Catalonia, or Provence; but most probably in Spain" (Erwin Panovsky, 1931, quoted by Ethel Lilian Voynich)
- highly abbreviated Latin (Joseph Martin Feely, 1943)
- written in England in medieval English (Leonell Strong, 1945)[11]
- a text in a synthetic language (William F. Friedman, 1940s?)
- "… the basis was more likely to be a very primitive form of synthetic universal language" (Brig. John R. Tiltman, 1951)[12]
- "German … der Musdel on f66r and so nimm geiss milch on f116v …" (Richard Salomon, prior to 1954)
- "… it was written … somewhere in Germany" (Erwin Panovsky, having second thoughts in 1954)
- "of Germanic origin" (Charles Singer, 1957)[13]
- "of the sixteenth century. of South German work and possibly related to Prague and John Dee." Czech, Polish or some other East-Central European language should be considered to underlie the text. (Charles Singer, 1957?)[14]
- "The country of origin is definitely European; it might be England, France, Italy, or what is now Germany … the text is based upon a written language that is probably Latin … but may be medieval English, French, Italian, or Teutonic." (Elizebeth Friedman, 1962)[15]
- "Italy is as likely a place of origin as any other country of Europe" (Hellmut Lehmann-Haupt, 1963)[16]
- a forgery, consisting entirely of the Arabic numerals one through nine (Robert S. Brumbaugh, 1975)[17]

- a “hitherto unknown North Germanic dialect” (James R. Child, 1976)
- a series of letters written in Ukrainian (John Stojko, 1978)
- a creole mainly based on Flemish (Leo Levitov, 1987)
- originating from Milan (Nick Pelling, 2006)
- associated with the Aztec culture in Meso-America (Arthur Tucker and R. Talbert, 2013)[18]
- a phonetic form of medieval Turkic (Ahmet Ardıç, 2018)
- and most recently, Hebrew (Rainer Hannig, 2020)[19]

Most of these propositions were the work of experts in their respective fields (though not all their professions were related to literature, linguistics, art, or history). Again, we see a parallel with the story of D. B. Cooper. In the case of Flight 305, each of the alleged suspects was supposed to resemble the artist’s impressions that the FBI had published in the newspapers; but the suspects did not at all resemble each other.

We can observe also that the Voynich manuscript lends itself to an obsession with minutiae. Not a few researchers have agonized, or at least perspired, over a single illustration, or a detail of an illustration, or a glyph that occurs only once in the whole manuscript. Here is an example, relating to the first page of the manuscript:

> Paragraphs 2 and 3 begin with big “weirdo” doodles, very unlike normal [sic] Voynichese letters. The first (EVA @252) looks like a capital K lying on it[s] side, with the vertical bar at the bottom. The second (EVA @253) looks like the first, with an extra squiggly line rising from between the two “horns.” The two symbols are drawn or painted with flaring strokes ending in swallowtail serifs.[20]

All of this is accurate, but where can two unique symbols lead us?

Like the “Cooper Vortex” (a term which some researchers have used to denote obsession with the case of D. B. Cooper), the manuscript has spawned “The Curse of the Voynich.” Like the inhabitants of the Cooper

Vortex, some of the researchers of the Voynich manuscript have lost their way in a dark forest, reminding us of Dante Alighieri in the opening lines of *La Divina Commedia*.

Finally, we turn to the characters with their inscrutable shapes and flourishes, whether lying mutely on the page or jostling with the illustrations that they explain or narrate (or so we are invited to believe). A few of them resemble Latin letters, Arabic numerals, or certain alchemical signs; the majority resemble no letter or character in any known extant or historic language. They require a systematic and objective assessment; this will follow in another chapter.

POSTSCRIPT: TOUCHING THE ELEPHANT

There is a well-known parable, that originated in what is now India in the fifth century BC, or probably earlier. It concerns a group of blind men who, for the first time in their lives, encounter an elephant. Each of them, by touching a part of the animal, attempts to relate it to something that is within his experience.

John Godfrey Saxe authored a poem on this story. It begins as follows:

> It was six men of Indostan
> To learning much inclined,
> Who went to see the Elephant
> (Though all of them were blind),
> That each by observation
> Might satisfy his mind.[21]

One blind man touches the trunk and declares that the object is a snake. Another feels the ear and says it is a fan. A third grasps the tail and is convinced that it is a rope. And so on … In some versions of the story, a

fight breaks out because none of the blind men can conceive that the object has all the properties that they separately perceived.

The history of research on the Voynich manuscript recalls this parable. The manuscript is to us what the elephant was to the blind men: it resembles nothing in our experience. One blind man pokes the manuscript in the region of its opening line and declares that the object is a document in Hebrew. Another runs his sightless fingers over the last page and concludes that it is German. A third touches the labels on the "pharmaceutical" drawings and imagines anagrams in Latin. A fourth detects symbols of magic or alchemy …

To this author, the Voynich manuscript is an integrated whole. It is not a Frankenstein collection of isolated parts. It was created by a producer or producers based on a vision or design. That person, or those persons, delegated the production to a team of scribes, based on a set of instructions. Those instructions had to be such that the scribes could understand and execute them without continuous recourse to the producers. Therefore, the instructions were simple.

The author did not tell one scribe to make a snake, another to make a fan, and another to make a rope. Metaphorically, he (or she, or they) told the scribes to make an elephant.

The Mandelbrot set (of which, more later) is a mathematical object of infinite complexity, but the equations which specify it can be written in a few lines. Even if we cannot translate the Voynich manuscript, perhaps we can re-create the author's instructions.

POSTSCRIPT: PAREIDOLIA

Scientists, especially behavioral scientists, use the term pareidolia to describe the human tendency to see patterns, and therefore meaning, in random images. Pareidolia has been recognized since antiquity and has been the subject of commentary by authors and artists as diverse as Leonardo da Vinci and William Shakespeare.

Pareidolia was at the source of at least two colorful theories that posited the existence of ancient civilizations on the planet Mars. One such theory

was that of the American astronomer Percival Lowell (1855–1916). Some observers in the nineteenth century, including the Italian astronomer Giovanni Schiaparelli, had recorded features on Mars that seemed to be linear on a scale of hundreds of kilometers. Lowell was probably the first to propound the idea that these features were irrigation canals, built by an intelligent race for their own survival. Modern space missions have found no such structures, nor even large-scale features that follow straight lines.

A recent and well-known instance of pareidolia led to the phenomenon of "the Face on Mars."

In 1976, during a routine mission, the American spacecraft Viking I took a photograph of a mesa, about 2 kilometers long, in the Cydonia region of Mars. The mesa, imaged at a low sun angle, seemed to resemble a male human face, gazing into space. NASA innocently published the image, with the comment that it was a "trick of light or shadow." They had not reckoned with the human love of mystery. Another image, taken from a different angle, was discovered by researchers. Soon the "Face on Mars" was the subject of books, monographs, and scholarly seminars, and even appeared in a movie.[22]

The "Face." we learned, was symmetrical, grandly proportioned, and had the majestic features of a king or emperor. It was not alone. Nearby were the remains of ancient structures, monoliths, colossal walls, even pyramids, clearly built by a long-lost civilization.

For many years, the mission controllers at NASA resisted demands for another photographic targeting of the "Face." Finally, they relented. In 2001, the Mars Global Surveyor took the first new image of the object, at a much higher resolution than that of the Viking. It was revealed to be an eroded mesa with a pleasing symmetry, and certainly with gulleys and hollows that conveyed elements of a human face. Whether that is the end of the story, the reader may decide. This author is content for the mesa to be the product of erosion, by wind or by water, and not the work of ancient Martians, however much we would like it to be so.

Likewise, determined researchers of the Voynich manuscript can find, within its cryptic and inscrutable pages, that which they wish to find.

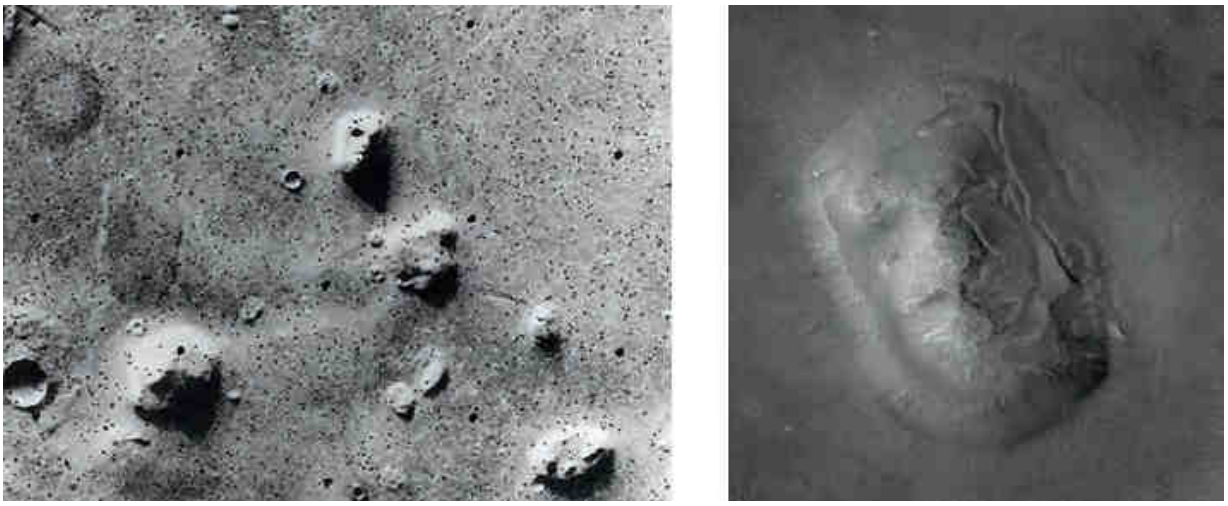


Figure 2.10. (*Left*) Part of the Cydonia region of Mars, including the “Face,” taken July 25, 1976, from Viking 1 Orbiter, Viking CD-ROM Volume 10, image F035A72. (*Right*) Mars Global Surveyor image of the “Face,” 2001. *NASA, JPL, Malin Space Science Systems*

CHAPTER 3

THE GLYPHS

At the heart of the Voynich mystery is the set of calligraphic shapes that have the appearance of text. Many researchers have described these shapes as “letters” or “characters”; and strings of such shapes, when delimited by spaces or by line breaks, they have called “words.”

In the sentence that you have just read, there were twenty-seven words. Whenever you saw a space, you took it to denote the end of a word. When you saw the quotation marks, the semi-colon, the comma, or the period, you understood that that these were delimiters of ideas, but they were not words or parts of words.

In the Voynich manuscript, there is nothing resembling punctuation. The shapes are separated by spaces of variable width, on a spectrum from zero (where consecutive shapes seem to touch) to a maximum of about 5 millimeters. The convention in Voynich research has been to regard spaces of maximum width as “word” breaks, and spaces of intermediate width as possible “word” breaks or alternatively as gaps within “words.”

The nomenclature of Voynich “letters” and “words” embodies an assumption that each shape corresponds uniquely to a letter as it is conventionally understood in natural human language. This assumption implies that each shape, if we could only map it to a letter of the Latin alphabet (or any other alphabet), could be pronounced by human lips and tongues; that strings of these shapes could form words with meaning; and that sequences of such words could form sentences that we could hope to understand.

This author prefers the term “glyphs.” Following the principle of Occam’s razor, we would like to avoid unnecessary assumptions. The glyphs may, or may not, represent letters in any natural language; if they do, they may not do so uniquely; and even if they do, the resulting text may, or may not, have meaning.

As we shall see in a moment, some of the simple glyphs are almost identical to Latin letters or Arabic numerals; a very few of the more elaborate glyphs resemble medieval alchemical symbols. One or two of the glyphs look like the ornamental flourishes that have been observed in one other medieval text. But it is universally agreed that the Voynich script, as an integrated entity, has never been seen in any other document. It is legitimate to pose the question as to whether it is the script of some forgotten language, or alternatively a forgotten script of a known language. To this author, following Occam’s razor, these seem complex assumptions, requiring a language or a script, or both, to have existed in the fifteenth century or later, and to have vanished without trace. A simpler assumption would be that there was an underlying natural language, or languages, and that the script was invented for the sole purpose of the manuscript.

If we adopt this assumption (which is this author’s inclination), it follows that the scribes of the Voynich manuscript worked from source documents in languages which they knew, and that they were instructed to transcribe those documents to the invented script, using a system of mappings which was furnished by the author or producer. What those mappings might have been is largely the subject of this book.

We may reasonably propose that the scribes knew the source languages very well, and that they learned the mappings assiduously before they put quill to parchment. We can say this because one overriding quality of the Voynich text is fluency. In over 40,000 “words,” over 150,000 glyphs, and

probably over half a million quill strokes, there is no sign of hesitation or confusion. As the lyric goes, “I’ll know my song well before I start singing.”[1]

We could reasonably add that, whatever instructions the producer gave to the scribes, they were simple. If the scribes did not commit the instructions to memory, at least they needed no more than a set of reference charts that we can imagine pinned to the walls, or spread on the tables, where the scribes were at work. Again, we can say this because the manuscript does not often give a sign that a scribe stopped work (to ask for advice or instructions?), allowed the ink to dry and returned with a fresh quill.

LANGUAGES

Among researchers of the Voynich manuscript, there is universal agreement that the glyphs were written from left to right, and from the top to the bottom of the page. There are 4,129 lines of horizontal text; in 3,885 lines, the glyphs are left-justified.[2] There is usually little, or no space left at the top of the page; there is often a generous blank area below the text. In this respect, the glyph sequences resemble European writing, and invite the hypothesis that they are based on a European language or languages; the researcher could be forgiven for excluding languages that are written from right to left, such as Arabic or Hebrew.

Before the widespread use of computers, any sharing of information on the glyphs had to involve manual copying of the shapes that were seen on the page. As we recall from our prologue, in the seventeenth-century Georgius Barschius did so in his attempts to interest Athanasius Kircher in a search for their meaning. Three hundred years later Prescott Currier, although he had access to computers, nevertheless had present his handwritten renditions of the glyphs for Mary D’Imperio’s Voynich symposium in Washington, DC.

Subsequently, several attempts were made to map the Voynich glyphs to the Latin letters, Arabic numerals, punctuation marks and special characters that appear on a modern American or European keyboard. To be clear: these

attempts were not, and were not intended to be, a translation of the manuscript. There was no implication that the mapping resulted in a pronounceable text. The intent was simply to make the glyphs computer-readable and thereby open the door to statistical analysis of their frequencies and other properties.

CURSIVE … OR NOT

The process of making the Voynich manuscript machine-readable necessarily involved assumptions about what constituted a glyph. In other words, the transcribers had to make decisions about where a glyph began and ended.

This decision involved an assumption about the extent, if any, to which the script of the manuscript was cursive; that is, whether consecutive characters were sometimes joined to each other.

Most European languages, at least if written in Latin or Cyrillic script, have a cursive form for handwriting and a noncursive form for printing. An exception is the Greek language, which is rarely written in a cursive form. If therefore, with no knowledge of European languages, we saw a printed text in any such language or even a handwritten text in Greek, we could reasonably assume that a space told us where one letter or character ended, and another began.

There are, however, languages which are only partially cursive. An example is Arabic. The basic Arabic alphabet has twenty-eight letters, of which twenty-two are normally written cursively. For example, the three Arabic letters ك (kaf), ت (ta) and ب (ba) when written cursively form the word كتب (kataba) or "he wrote." Only if we read Arabic can we know that this shape consists of three letters and not two, or even one. However, there are six Arabic letters which are never connected to their neighbors: ا (alif), د (dal), ذ (dhal), ر (ra), ز (zayn), and و (waw). Here is a word consisting entirely of these letters: إدواردز or "Edwards"; in this case we can reasonably assume that we see seven letters.

In the Voynich manuscript, there are some very common shapes which appear to stand alone with no cursive connections to their neighbors: they

include , , , , and . It has universally been assumed that each of these shapes is a glyph; and this is a reasonable assumption. But what are we to make of shapes such as , , or ? It has been widely (but not universally) assumed that each of these is a glyph. It is clear that each shape consists of several strokes of the quill, and that the quill has been lifted after every stroke. But does the lifting of the quill imply the continuation of a single glyph, or the beginning of a new one?

Even the common shape , which never shows a sign of a lifted quill, could be interpreted as a combination of two shapes, namely and .

If we had no knowledge of the English language, and we saw a handwritten letter “w,” we might suppose it to be one letter, or we might see it as two letters each being “v” (as it originally was). If we saw the letter “m,” we might hesitate between treating it as a single letter or as a concatenation of “i” and “n,” or of “r” and “n.”

A conservative mapping of the Voynich glyphs would call, in any cases of uncertainty, for breaking down the glyphs into fragments, each with a minimal number of strokes of the quill and leaving it to subsequent analysts to recombine the fragments as they thought fit.

However, most machine-readable transcriptions of the Voynich manuscript have assumed that the script is not cursive; on this assumption, lifting the quill does not denote a new glyph. The transcribers needed to see clear blue sky, so to speak, between quill strokes before a new glyph was identified.

Even so, any analyst who has recourse to a computer-readable version of the Voynich manuscript must bear in mind that any such document was created by a fallible human being. We must imagine that that person read the manuscript, page by page, line by line, and glyph by glyph, in all its daunting complexity and opacity, and made decisions of the following nature:

- This shape is a glyph, and I will assign it such and such a symbol on the keyboard.
- This other shape is very similar, but it is not the same glyph, and I will give it another assignment.

- This other shape is made up of many quill strokes but nevertheless I will consider it a single glyph.
- This other shape contains several glyphs that I already assigned, but nevertheless I will give it a separate assignment.
- This other shape could be a glyph that I already assigned, but it is wider, or taller, or has some extra curlicues: I will call it a "rare glyph."
- and so on …

Computer readability is a double-edged sword. It allows us to search efficiently for patterns, repetitions, and frequencies, but we lose sight of the original glyphs in all their elegance and mystery. We can become lost in the statistics, and we can forget that the transcriptions are human interpretations.

V101

In order, with all these reservations, to apply modern computing power to the Voynich manuscript, this author started out with the mappings of Glen Claston,[3] who created the "Voynich 101" or "v101" alphabet, more precisely the "v101 keyboard assignments."

The v101 alphabet identifies seventy-one glyphs as the core of the Voynich glyph set. Claston gave each of these glyphs an assignment on the English-language keyboard, using all twenty-six lowercase letters, twenty-five uppercase letters (all except "O"), the nine Arabic numerals from 1 to 9, and eleven punctuation marks or special characters such as "!" and "#." For wide spaces between strings of glyphs, v101 uses the period {.}, and for narrow spaces where it is unclear whether there is one string or two, v101 uses the term "uncertain space," represented by the comma {,}.

Figure 3.1. Timothy Rayhel (Glen Claston). *Nick Pelling*

The simultaneous use of lowercase and uppercase Latin letters is not problematical if the keyboard assignments are imported to a document format such as Microsoft Word, which can recognise cases. If the import is to Microsoft Excel, which has some functions that are not case sensitive, it requires extra care to avoid confusion between, for example, the glyphs represented by {c} and {C}, and those represented by {k} and {K}, all four of which are relatively common.

Figure 3.2. Glen Claston's identification of the basic Voynich glyph set, and his corresponding keyboard assignments in the v101 alphabet. *Glen Claston, additional graphics by the author*

Claston's v101 keyboard assignments are summarized in figure 3.2.

TOO MANY GLYPHS

We may surely be permitted to observe here, firstly, that seventy-one glyphs are far too many to correspond to the alphabet of a natural phonetic language (as are all Indo-European and Semitic languages), even if we allow ten glyphs to represent Arabic numerals; and far too few to map to a symbolic language such as Chinese, in which characters represent words rather than sounds.

If our working hypothesis is that the underlying languages are European, then we have too many glyphs, and some of the keyboard assignments must be duplicates, or represent variations in handwriting but do not represent distinct glyphs or are concatenations of several glyphs.

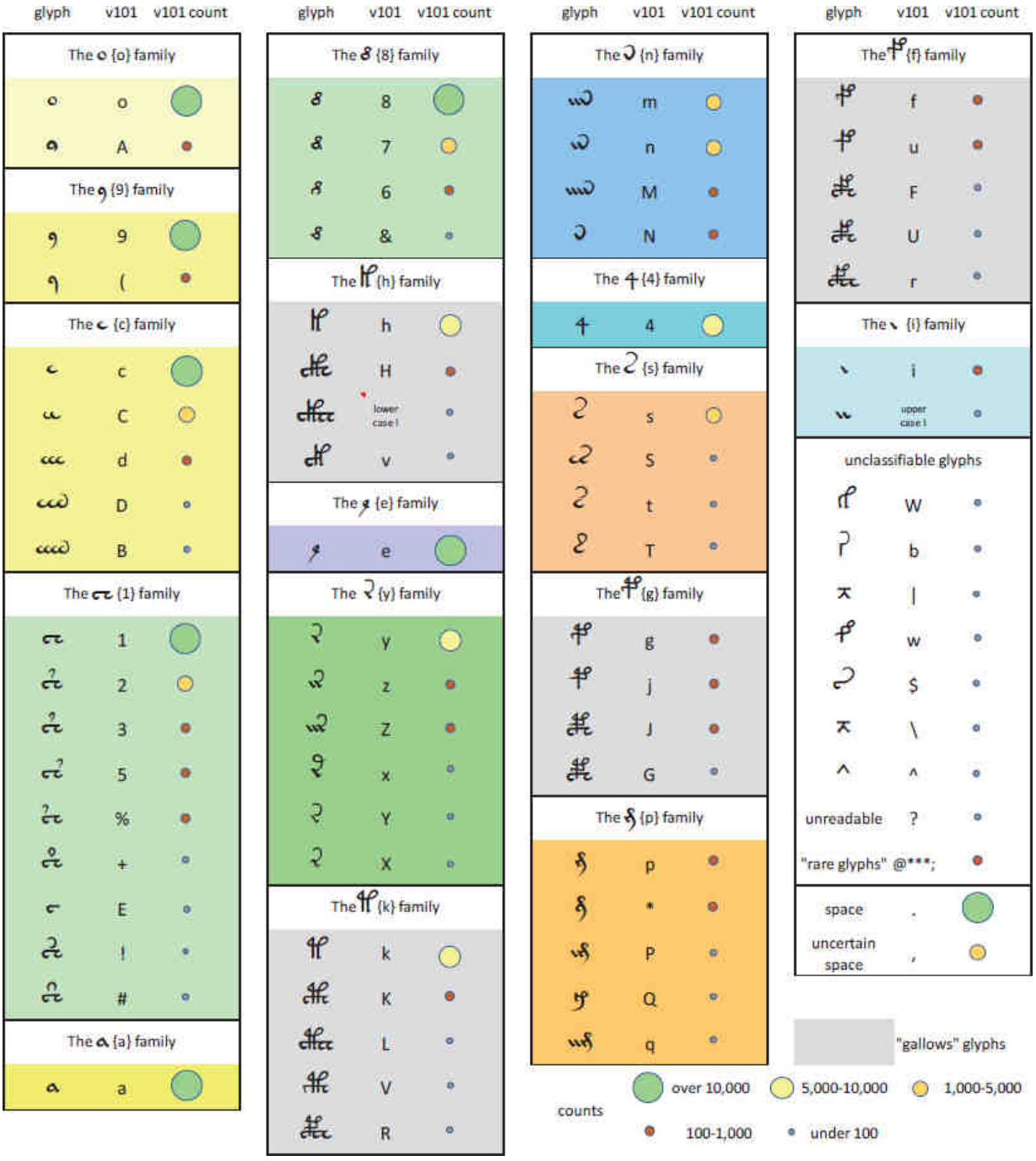
Secondly, although it is not obvious from looking at the v101 alphabet, some of these glyphs are extremely infrequent. In the English language, we probably sense that "e" is common, and "z" is not; and this is true—in fact "e" is about 210 times more common than "z." But among the seventy-one glyphs in the basic v101 alphabet, the most common glyph occurs 4,124 times more often than the least common. The glyph , which receives the keyboard assignment {B}, occurs only six times. If we are to use the v101 alphabet, we need to be aware of which glyphs we need to keep in mind, and which ones we can forget. To this end, Table 3.1 summarizes the v101 glyphs in terms of their frequencies of occurrence in the manuscript.

Table 3.1 broadly illustrates the counts of the glyphs which appear in the basic v101 set. The seventy-one glyphs can be grouped into five categories, in terms of their frequency of occurrence in the manuscript, as follows:

- Seven glyphs each occur over 10,000 times in the manuscript.
- Four glyphs each occur between 5,000 and 10,000 times.

- Six glyphs each occur between 1,000 and 5,000 times.
- Twenty-one glyphs each occur between 100 and 1,000 times.
- The remaining thirty-three glyphs occur less than 100 times each, and each accounts for less than 0.05 percent of the text. These glyphs are rarer than the letter “z” in the English language, which has a frequency of 0.06 percent.

Table 3.1: Glen Claston’s basic glyph set and his corresponding keyboard assignments in the v101 “alphabet”



Source: Zandbergen, Rene. *Text Analysis - Transliteration of the Text, Table 6: the v101 basic character set*; glyph relationships by author; glyph counts by author based on v101 transliteration

On the basis of these assignments (and their underlying assumptions about what is and what is not a glyph), the Voynich manuscript contains 158,947 glyphs.[4] If we assume that a space, or an uncertain space, or a line

break delimits a word, the Voynich manuscript contains 40,708 words. (Different definitions of a glyph would yield different counts of glyphs; different definitions of spaces would yield different counts of words.)

RESEMBLANCES

We may firstly remark that not all of the glyphs are completely alien to us. Some of them bear a resemblance to letters or numerals which exist in modern European languages. For example:

- ɑ {a} resembles the Latin letter “a”
- ɔ {c} resembles the Latin letter “c”
- o {o} and ɑ {A} resemble the Latin letter “o”
- † {4} is similar to the Arabic numeral 4
- ȣ {8} and its less common variants ȣ {6}, ȣ {7} and ȣ {&} are similar to the Arabic numeral 8
- ọ {9} and ọ {(} are similar to the Arabic numeral 9.

No doubt these resemblances underlay some of Glen Claston’s keyboard assignments. However, there is no implication that these glyphs stand for the letters or numerals that they resemble. With regard to the glyphs that resemble numerals, there is indeed no implication that there are any numerals in the Voynich manuscript.

Some of the glyphs appear to occur quite frequently in doubled or even tripled forms (although in the v101 alphabet, these are usually treated as parts of other glyphs). For example:

- The glyph ıı, resembling a double ı {i}, has its own assignment {I} but it also looks like part of the glyphs ıın {m}, ıȣ {P} and ıȓ {z}.

- What looks like a triple i {i} can be seen hiding in the glyphs ᔑᔑᔑ {M}, ᔑᔑᔕ {q} and ᔑᔑᔖ {Z}.
- The glyph ᑢᑢ, which resembles a doubled ᑢ{c}, has its own assignment {C} but could also be viewed as part of the glyph ᑢᑢᑢ {d}, and of the rare ᑢᒋ {S}, ᑢᑢᒋ {D} and ᑢᑢᑢᒋ {B}.

Claston's assignments suggest to us what an alien, never having seen the English language, might have done with repeated English letters such as "ee," "oo" or "nn"; he, she or it might have assumed that these were distinct letters.

Since tripled letters are very rare in European languages, it might be worth our while to consider at some point whether these glyphs, and their doublets and triplets, are in any way related to Roman numerals, for example:

- I Roman 1
- II Roman 2
- III Roman 3
- C Roman 100
- CC Roman 200
- CCC Roman 300

DUPLICATES

Without claiming any expertise in handwriting analysis, this author is tempted to hypothesize that the basic v101 glyph set has many duplicates, including at least the following:

- keyboard assignments ᓀ {2}, ᓂ {3}, ᓄ {5}, ᓇ {!} and ᓈ {#} probably represent the same glyph
- ᔓ {6}, ᔕ {7}, ᔖ {8} and ᔗ {&} probably represent the same glyph

- ৭ {9} and ৭ {(} probably represent the same glyph
- ৲ {a} and ৲ {A} are probably the same glyph
- ৱ {s} and ৱ {t}, and possibly also ৪ {T}, are the same glyph
- ৸ {x}, ৲ {X}, ৲ {y} and ৲ {Y} are probably the same glyph.

If this hypothesis bears scrutiny, then the basic glyph set could be reduced from seventy-one to fifty-seven. That is still too many for a natural language, but it is a start.

The user of the v101 alphabet, especially if he or she is not referring simultaneously to the original glyphs, is tempted into a reading which is imposed by the alphabet and allows no alternatives. This author can see at least fourteen cases where the v101 alphabet could have permitted an alternative interpretation if consecutive strokes of the quill were allowed to be perceived as distinct glyphs. These are listed in Table 3.2.

Table 3.2: Selected possible alternatives to some common keyboard assignments in the v101 alphabet

glyph family	glyph	v101	v101 count	alternatives					
				glyph	v101	glyphs	v101	glyphs	v101
glyphs related to ৫ {c}	৫৫	C	2,844	৫৫	cc				
	৫৫৫	d	336	৫৫৫	ccc	৫৫৫	cC	৫৫৫	Cc
glyphs related to ৷ {i}	৷৷	I	98	৷৷	ii				
glyphs related to ৩ {N}	৷৷৩	m	4,112	৷৷৩	iiN	৷৷৩	IN		
	৷৩	n	1,770	৷৩	iN				
	৷৷৷৩	M	174	৷৷৷৩	iiiN	৷৷৷৩	iIN	৷৷৷৩	IiN
glyphs related to ৺ {p}	৷৺	P	989	৷৺	ip				
glyphs related to ৲ {y}	৷৲	z	511	৷৲	iy				
	৷৷৲	Z	154	৷৷৲	iiy	৷৷৲	Iy		

Source: v101 keyboard assignments from Zandbergen, René: *Text Analysis - Transliteration of the Text*, Table 6: the v101 basic character set; alternatives by author.

If these alternatives have any validity, then we might be able to add one glyph and delete thirteen from the basic glyph set, thereby reducing the number from fifty-seven to forty-four, which is not too remote from the alphabet of a modern European language. Modern French, for example, has forty-two distinct letters, including the variants of vowels which are distinguished by diacritics.

Be that as it may, the v101 alphabet at least gives researchers a consistent database on which to perform statistical analyses and to search for meaning. In this book, the author has used the v101 keyboard assignments as a starting point and has diverged from them as and when it seemed necessary.

THE EXTRAORDINARY o {O} AND 9 {9}

The first observation that we must make about the glyphs is that there are two which occur with extraordinary frequencies, which are out of line with most European languages.

The glyph o {o} is by far the most common glyph. It occurs 25,043 times in the whole manuscript, and thereby accounts for 15.9 percent of the total glyph count (assuming that spaces and uncertain spaces do not count as glyphs). There are few European languages in which the most common letter is more frequent than this: the only ones that we have identified are Dutch (e = 19.3%), German (e = 16.9%), Norwegian (e = 16.6%), and Belarusian (a = 16.4%).

However, the glyph o {o} is stranger than this. If we treat spaces, uncertain spaces and line breaks as word breaks, there are 40,708 words in the Voynich manuscript. On this definition, o {o} occurs 8,794 times as the initial glyph of a Voynich "word"; in other words (no pun intended) it is the first glyph of 22.4 percent of all the Voynich "words."

As far as we can determine, there is no publicly accessible database of the frequencies of initial letters in European languages. To construct such a database would involve searching through large corpora of each of the

languages under consideration: which is a task far beyond the scope of this book. However, we are indebted to Peter Norvig for doing this exercise on the English language. Based on a corpus of nearly 744 billion words in English, distilled from Google Books, he has reported that the most common initial letter in English is "t," with a frequency of 16.0 percent.[5]

The Voynich glyph o {o} is a much more frequent initial character than that.

The second most common glyph, though a distant second to o {o}, is 9 {9}, which occurs 17,461 times and accounts for 10.7 percent of the total glyph count. That is not an unusual frequency for the most-common letters in European languages, for example, "e" in Albanian, Czech, Dutch, English, German, Norwegian, and all the Romance languages. What makes 9 {9} unusual is that it occurs 14,776 times as the final glyph of a Voynich "word"; that is, it is the final glyph of 37.7 percent of all the "words" in the manuscript.

Again, for comparison we may refer to Peter Norvig, who has calculated that the most common final letter in English is "e," with a frequency of 20.1 percent.[6] We would hazard a guess that if any European languages have a more common final letter than that, they would be Romance languages with their frequent word endings of "a," "e," "i" and "o." Later, we shall see that in Dante Alighieri's *La Divina Commedia*, the most common final letter of a line is "a," with a frequency of 30.7 percent. That comes close to matching the 9 {9} in the Voynich manuscript; but then, *La Divina Commedia* is poetry, where the lines typically have to rhyme.

We shall return to the glyphs o {o} and 9 {9} in later chapters. For now, the observation that we must make is that, if at all there is a mapping from Voynich glyphs to letters in Latin script, these two glyphs in particular (and indeed others) may not map one-to-one to letters. The mapping may depend on the position of the glyph relative to the word. For example, o {o} may have a different mapping depending on whether it is in the (frequent) initial position; or an internal position within the word; or a final position; or in the relatively few cases where it is a single-glyph word. Likewise, 9 {9} in its (frequent) final position may have a specific mapping; elsewhere, it may mean something else.

THE STORY OF ſo

Even more extraordinary than the o {o} and the 9 {9} is the shape that has universally been assumed to consist of two separate glyphs: ſo {4o}. This shape occurs 5,265 times in the manuscript.

The presumed glyph ſ {4} occurs 5,408 times, but in 5,265 cases it is part of ſo {4o}, and in only 143 cases does it seem to occur on its own.

We are tempted to infer the following:

- that ſo {4o} is a single glyph
- that ſ {4} is not a separate glyph; and that the instances where it seems to be separate are errors of transcription.
- that we can subtract 5,265 instances of o {o}, reducing its count to 19,778 or 12.6 percent of the total glyph count. Its frequency then starts to look more like that of a letter in a European language, although still probably the most common one: in other words, similar to "e" or "a."

But there is more to the story of ſo {4o}. It occurs 5,238 times as the beginning of a Voynich "word"; 27 times internally to a "word"; and never at the end of a "word."

THE GALLOWS GLYPHS

Researchers of the Voynich manuscript agree that there is a group of eighteen glyphs which are not only taller than most, but also have one or two vertical quill strokes, embellished with loops and curls. They are widely known as the "gallows glyphs." They are illustrated in Table 3.3.

By this author's calculation, there 19,872 "gallows" glyphs in the whole Voynich manuscript: or 12.5 percent of the total glyph count. Two of the

"gallows" glyphs are relatively common; {h} occurs 9,837 times (6.2 percent of the total glyph count) and {k} occurs 6,030 times (3.8 percent of the glyph count). None of the other sixteen "gallows" glyphs occurs more than 1,000 times or accounts for more than 1 percent of the glyphs in the manuscript.

Table 3.3: The "gallows glyphs" in Glen Claston's basic glyph set and the corresponding keyboard assignments "alphabet"

glyph family	"plain" variant			"pedestal" variant			"grand piano" variant		
	glyph	v101	v101 count	glyph	v101	v101 count	glyph	v101	v101 count
glyphs related to {h}		h	9919		H	866		lower case l	36
					v	23			
glyphs related to {k}		k	6118		K	918		L	42
					V	23			
glyphs related to {g}		g	879		G	89		R	11
		j	564		J	122			
glyphs related to {f}		f	233		F	35		r	12
		u	177		U	32			

Source: Zandbergen, Rene: *Text Analysis - Transliteration of the Text, Table 6: the v101 basic character set*; classification of variants by author; glyph counts by author, from v101 transliteration in Microsoft Word.

We have said that there are eighteen "gallows glyphs"; but this author will admit to seeing no substantive difference between {f} and {u}; between {F} and {U}; between {g} and {j}; and between {G} and {J}. Moreover, {v} looks like an unfinished {H}, and {V} resembles a hurried or mis-drawn {K}. A more conservative enumeration of the "gallows glyphs" might be twelve.

We may observe that the "gallows glyphs" have three variants:

- a “plain vanilla” variant (six glyphs), consisting of the vertical stroke or strokes and the curlicues;
- a “pedestal” variant” (eight glyphs), in which the plain glyph is superimposed on a , which seems like a lengthened version of the common glyph {1}
- a “grand piano” variant (four glyphs), in which the plain glyph is superimposed on a , which has no assignment in the v101 alphabet but resembles a combination of the glyphs {1} and {c}.

The “gallows” glyphs, despite their size, clearly do not correspond to uppercase letters as we know them in European languages; for there are no miniature versions with a comparable height to the “normal” glyphs.

Likewise, the “gallows” glyphs are not the initial glyphs of proper names; for they can occur at the beginning or the middle of a glyph string, or (though extremely rarely) at the end of a glyph string.[7]

Conceivably, the “gallows” glyphs could be equivalent to two-letter strings (digrams) in some natural language. If we want to pursue this hypothesis, we might look for the most-common digrams in selected European languages. If the Voynich manuscript were derived from Latin, then the most common “gallows” glyph, {h}, could be equivalent to the most common Latin digram “er”; likewise, {k} could be equivalent to “um.”[8] However, in this case we would have to ask why the parts of the “gallows” glyphs do not occur on their own. For example, although there are many instances of {h} and {k}, there are no instances of or or preceded or followed by other glyphs.

If the “gallows” glyphs have some specific function in the Voynich manuscript, no one has yet determined what that function is.

PUNCTUATION

We said earlier that there is no sign of punctuation. It is true that there is nothing resembling a period, a comma, a colon, a semi-colon, a question

mark, or an exclamation mark. There are no obvious quotation marks or parentheses. There are some shapes resembling apostrophes, but in all cases, they are linked to the glyph {2} and its variants, and not uniformly distributed throughout the text, as we would expect of an apostrophe.

Nevertheless, we believe that there are phenomena which have the same effect as punctuation. We shall elaborate in a later chapter.

THE SAN SAVINO DOCUMENT

We wrote in chapter 2 that the script of the Voynich manuscript had never been identified in any other document. That is true for the script as a whole; but shapes resembling some of the Voynich glyphs have appeared elsewhere. In the case of the "gallows" glyphs, there are at least two examples in a medieval document. The document is reproduced in Adriano Cappelli's *Lexicon Abbreviaturarum: Dizionario di Abbreviatture Latine ed Italiane*, first published in 1899. Cappelli introduced the document as follows:

> 1172, June 13. – Savino, abbot of the monastery of San Savino in Piacenza, grants to the miller Gerardo Albarola, for himself and his male heirs in perpetuity, a mill in the region of the said monastery.—Carolina script.—Original parchment, preserved in the State Archives of Parma, monastery of San Savino.[9]

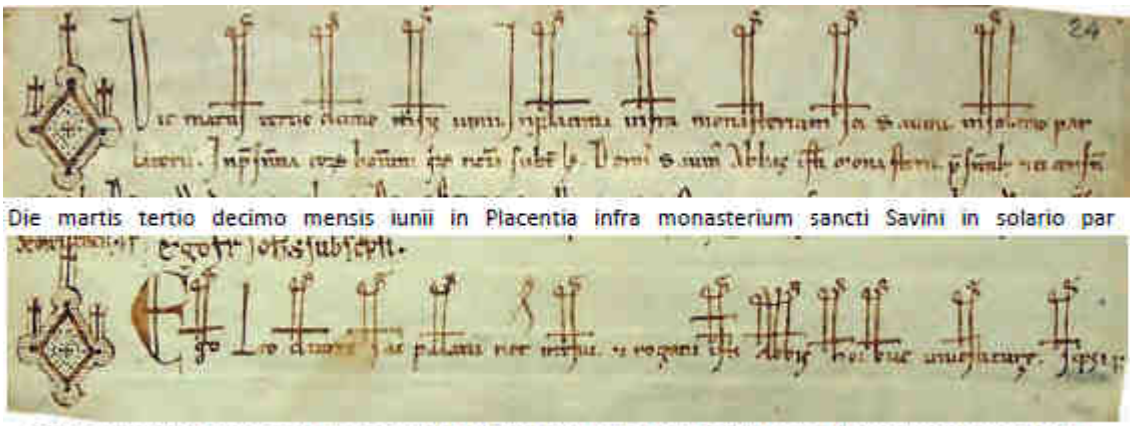

Die martis tertio decimo mensis iunii in Placentia infra monasterium sancti Savini in solario par
Ego Leo de Turre sacrae palatii notarius interfui et rogatu iamscripti Abbatis hoc breve investiturae scripsi

Figure 3.3. Document from St. Savino monastery dated June 13, 1172, first two lines and last line, from the document "Diplomatico, Atti privati, cassetto 6, doc. 373," by permission of Ministero della cultura, Archivio di Stato di Parma.

The first and last lines of the document, reproduced in figure 3.3, contain shapes which at first glance could be taken for glyphs from the Voynich manuscript.

The first line reads, in Cappelli's interpretation of the abbreviated Latin:

> die martis tertio decimo mensis iunii in Placentia infra monasterium sancti Savini in solario par||latorii."
>
> {on Tuesday, the thirteenth of June in Piacenza below the monastery of Saint Savino on the terrace of the audience-chamber.}[10]

In this line, eleven letters are embellished with vertical flourishes, of which eight have two curlicues and resemble the glyph {k} of the Voynich manuscript, although greatly elongated. There is even a suggestion of the "pedestal glyph" {K}, as shown in figure 3.4.

However, the flourishes do not seem to be linked to any specific Latin letter. If we rewrite the line with capitalization of the embellished letters, it becomes (as nearly as we can tell):

Die martiS tertio Decimo meNSis iunii In pLacentia inFra monaSterium Sancti savini in Solario par-latorii.

The looped flourish does not have to be on the initial letter of a word; it can be attached to one letter or two; and if a flourish has any preference for the letter on which it sits, it could be D, F, L, N, S, or T (but apparently it never sits on a vowel). There are two cases where the letters I and L are elongated vertically (but without loops).

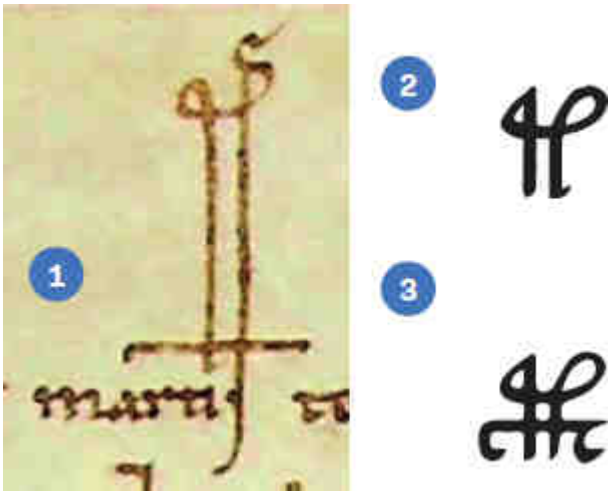


Figure 3.4. (1) Document from St. Savino monastery dated June 13, 1172, an example of the flourish; from the document "Diplomatico, Atti privati, cassetto 6, doc. 373" (by permission of Ministero della cultura, Archivio di Stato di Parma); (2) the Voynich glyph {k}; (3) the Voynich glyph {K}.

The last line of the document has thirteen flourishes of a similar nature; eleven of them resemble the Voynich glyph {k} or the {K}. Cappelli's transcription of this line, in which he interpreted the abbreviations, reads:

Ego Leo de Turre sacri palatii notarius interfui et rogatu iamscript Abbatis hoc breve investiturae scripsi

["I, Leo de Turre, notary of the sacred palace, was present and at the request of the said Abbot wrote this short investiture."][11]

Our own transcription, leaving the abbreviations as they are, and capitalizing the letters to which the flourishes are attached, reads:

> EGo Leo Dtur …Rae paLatíí not intRFui & rogatu iMSti ABBTis Hoc Bve inveSTiturae SCRipsi.

We note that the superscript , whereby "notarius" is abbreviated to "not," has more than a passing resemblance to the glyph ꝺ {*}, which is nearly always in the final position of a Voynich "word."

The resemblance of the vertical flourishes to two of the Voynich "gallows" glyphs is intriguing, but they are not a script, and they have no apparent meaning; they are just decorations of the Latin letters. However, the abbreviation of "notarius" is a possible clue to what the Voynich glyphs are all about. This we shall pursue in another chapter.

Regarding Cappelli's identification of the script as "Carolina," it seems probable that his reference was to Carolingian minuscule or Caroline minuscule, which was a standard script for sacred and profane texts in central Europe from the ninth to the thirteenth centuries. Even so, the flourishes on the document from the St. Savino monastery seem to be unique. We might have hoped for a key to the Voynich manuscript, but what we have found seems to be just a coincidence.

However, we may conjecture that the authors of the Voynich manuscript had seen, and even had been inspired by, the San Savino document or one resembling it. If so, we could suspect that the "gallows glyphs" are not letters at all, but merely decorative flourishes in the San Savino style. In that case, without changing the meaning of the Voynich text, all 17,705 of the "plain gallows" glyphs could be eliminated; the 2,078 instances of "pedestal" glyphs could be reduced to a simple ᖁ {1}; and the eighty-nine instances of "grand piano" glyphs could be reduced to a ᖵ which might be construed as ᖁᖁ {E1} or ᖁᖀ {1c}, or might be a single glyph in its own right.

If we may take this hypothesis a stage further: since the San Savino document never added flourishes to vowels, it suggests to us that ᖁ {1}, and ᖵ represent consonants.

A revised Voynich alphabet, in which the "gallows" glyphs are essentially eliminated and the residual ⫏ {1} ⫐ and are mapped to appropriate consonants, could be a productive area of research, for whomsoever is so inclined.

GESSMANN AND SOMMERHOFF

Many researchers of the Voynich manuscript, notably Mary D'Imperio (of whom, more in chapter 6) have identified resemblances of some of the Voynich glyphs to certain symbols used in medieval alchemy. This author is using the term "alchemy" to mean, not magic or pseudoscience, but the medieval form of chemistry, in which philosophers and applied scientists sought to understand the relationships between different materials, and the possibilities of transforming one material into another.

D'Imperio took as her primary source the second edition of Gustav Wilhelm Gessman's book *Die Geheimsymbole Der Alchymie, Arzneikunde Und Astrologie Des Mittelalters* [*The secret symbols of alchemy, pharmacy, and astrology of the Middle Ages*], published in Berlin in 1922. Although Gessmann referred to these symbols as secret, he nevertheless provided interpretations in English, French, German, Italian and Latin for all but a few of the symbols that he included in his book. We could infer that he had access to reference materials which translated the symbols into European languages.

Gessmann did not make any claim that his book was an exhaustive compilation:

> … to assemble the thousands and thousands of composite symbolic representations which were created in the Middle Ages would be the work of a human life.[12]

However, he did write a brief (twenty-page) history of alchemy in Europe from the year 115 AD to the nineteenth century (when apparently, some forms of alchemy were still being practiced). Extraordinarily, this narrative intersected with that of Wilfrid Voynich. We recall from our

prologue that Voynich nominated the English adventurer John Dee as the purveyor of the "Roger Bacon cipher manuscript" to the emperor Rudolf II. In his presentation of 1921 in Philadelphia (one year before the second edition of *Die Geheimsymbole*), Voynich wrote:

> … In the collection of Bacon manuscripts, which unquestionably came into his [Dee's] possession as early as 1547, he found the cipher manuscript. The sequence of events which suggest themselves is that having failed to decode it, he carried the manuscript to Prague, where he parted with it as a "present" to Emperor Rudolph. He can only have presented it to the emperor between 1584 and 1588, during which years he made several visits to Prague.[13]

Gessmann had more on this story:

> Two more important alchemists of this time [the sixteenth century], who enjoyed the favor of Emperor Rudolf, were the Englishmen Edward Kelly and Dr. John Dee. These two, who presented themselves to the emperor as manufacturers of the [alchemical] tincture and were actually supposed to have made changes before the emperor's eyes, but [were] unable to prove their claims, were repeatedly incarcerated and finally perished miserably.[14]

Gessmann's narrative was essentially true in respect of Edward Kelley, but not of Dee. In Bohemia the two Englishmen were indeed collaborators, but Dee returned from Prague to England in 1589, and never again set foot in Bohemia. One wonders whether Voynich had read Gessmann's first edition (which had been published in 1906) and wove it into his story of John Dee as the merchant, not of alchemical miracles but of mysterious manuscripts. There is no document that supports Voynich's hypothesis, so we can make of his story what we will.

However, our interest in Gessmann's work is not his story of palace intrigues in Bohemia, but the alchemical symbols themselves. D'Imperio wrote in *An Elegant Enigma* as follows:

> Gessmann (1922) presents a large collection of the symbols and code words used by medieval alchemists and other scholars and philosophers…. some … appear similar to certain Vovnich script characters. It was apparently common practice for alchemists to employ these symbols, interspersed in Latin text, as a sort of secret shorthand for alchemical products and processes. While a few of these signs are somewhat similar to Voynich symbols. most of them are not, and they offer disappointingly little help in our task [of decipherment].[15]

It is indeed true that in the 120 pages of Gessmann's book which he titled *Tafeln der Grundsymbole* ("Tables of the Basic Symbols"), there are over 1,000 separate symbols, of which, by this author's count, only forty-six are similar to some of the Voynich glyphs. Among these forty-six, many are very similar to each other. Taking the correspondence in the opposite sense, this author can identify only fourteen Voynich glyphs, as defined in the v101 alphabet, which seem similar to symbols in Gessmann's book. They are as shown in Table 3.4.

Some of these similarities could be dismissed as coincidence. It would not be surprising if a medieval scientist used symbols similar to the Arabic numerals 4, 8, and 9 and the Latin letters "a," "c," "m," and "o." Furthermore, Gessmann's tables make it clear that such symbols were not shared among alchemists, and therefore, we suppose, could have multiple meanings, depending on the individual predilections of the alchemist. For example, the symbol resembling an "8" could mean aqua vitae, bull's pizzle, gold, gold-spirit (whatever that might have been), hematite, water, or white oxide of arsenic; or it could be a generic symbol for mineral salt or sublimate.

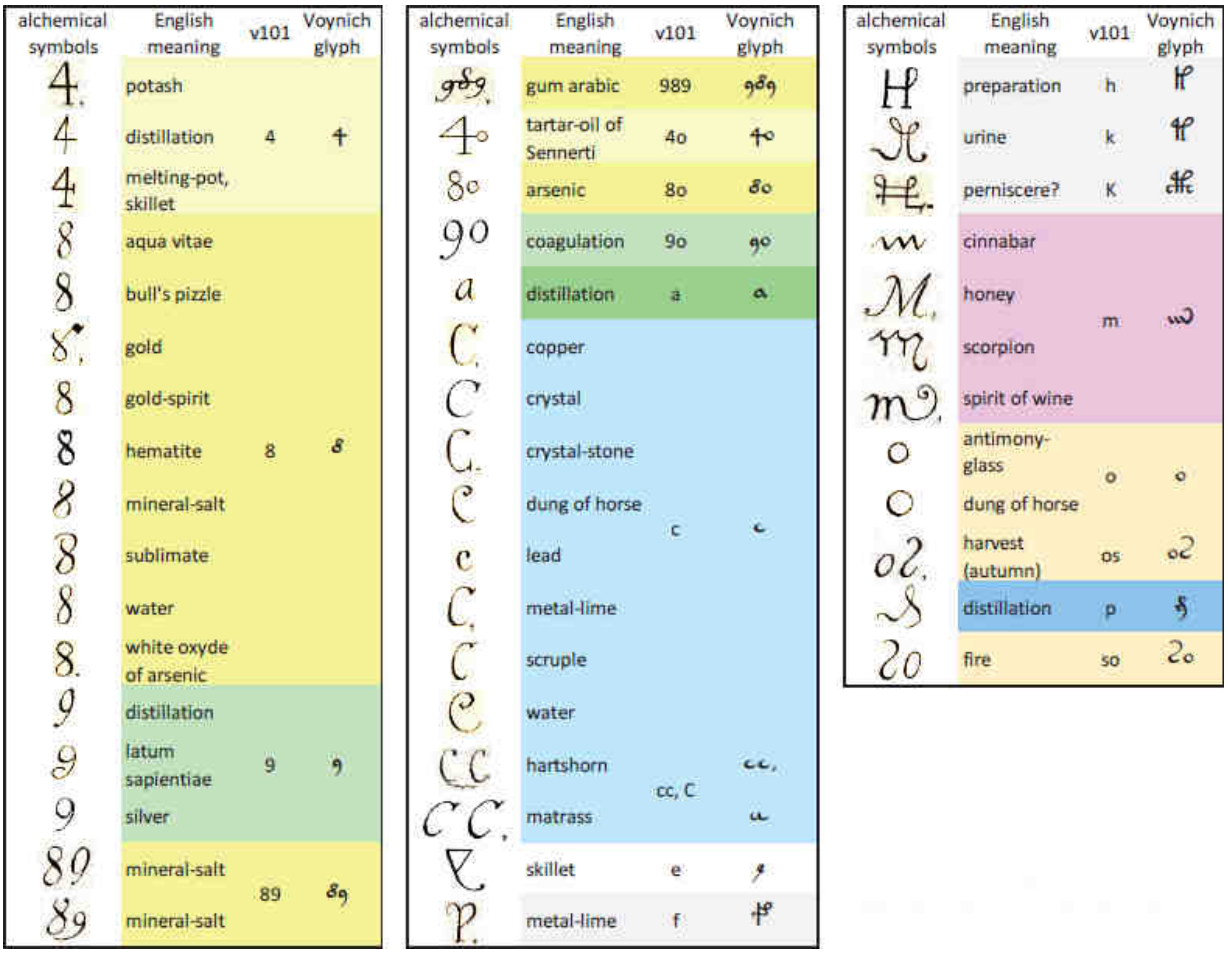
Nevertheless, among the resemblances there are some surprises.

For instance, the alchemical symbol bears an uncanny resemblance to the glyph string , which occurs 5,265 times in the Voynich manuscript. We recall that {4} rarely occurs without the {o} attached, which, in this author's view, means that the glyph string is probably not two glyphs, but one. Gessmann interpreted this symbol as "Weinsteinöl, Sennerts" (tartar oil of Sennert).[16]

Three other symbols in the Gessmann compendium are intriguing, though in these cases Gessmann gave no hint of their origins or the names of the alchemists who might have used them:

- which Gessmann read as “preparation”[17] and which resembles the glyph {h}, which occurs 9,837 times in the Voynich manuscript
- which Gessmann interpreted as “urine”[18] and which has a passing resemblance to the Voynich {k} (6,030 occurrences)
- for which Gessman’s tentative reading was “perniscere,”[19] probably a misspelling of the Latin “permiscere” (“thorough mixing”); here we see a resemblance to the Voynich “pedestal glyph” {K} (906 occurrences).

Table 3.4. Selected alchemical symbols identified by G. W. Gessmann, and morphological resemblances to Voynich glyphs



alchemical symbols	English meaning	v101	Voynich glyph
	potash		
	distillation	4	
	melting-pot, skillet		
	aqua vitae		
	bull's pizzle		
	gold		
	gold-spirit		
	hematite	8	
	mineral-salt		
	sublimate		
	water		
	white oxyde of arsenic		
	distillation		
	latum sapientiae	9	
	silver		
	mineral-salt	89	
	mineral-salt		

alchemical symbols	English meaning	v101	Voynich glyph
	gum arabic	989	
	tartar-oil of Sennerti	4o	
	arsenic	8o	
	coagulation	9o	
	distillation	a	
	copper		
	crystal		
	crystal-stone		
	dung of horse	c	
	lead		
	metal-lime		
	scruple		
	water		
	hartshorn	cc, C	
	matrass		
	skillet	e	
	metal-lime	f	

alchemical symbols	English meaning	v101	Voynich glyph
	preparation	h	
	urine	k	
	perniscere?	K	
	cinnabar		
	honey	m	
	scorpion		
	spirit of wine		
	antimony-glass	o	
	dung of horse		
	harvest (autumn)	os	
	distillation	p	
	fire	so	

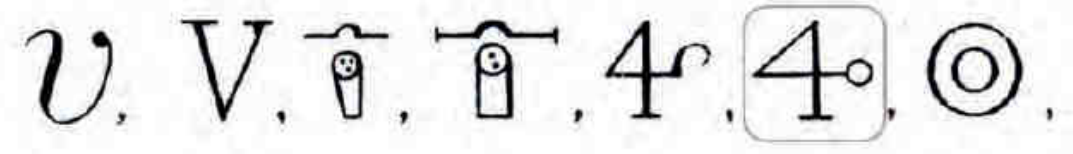
Source: alchemical symbols from Gessmann, G. W.: *Die Geheimsymbole Der Alchymie, Arzneikunde Und Astrologie Des Mittelalters*, pp. III-CIX; resemblances to Voynich glyph identified by author

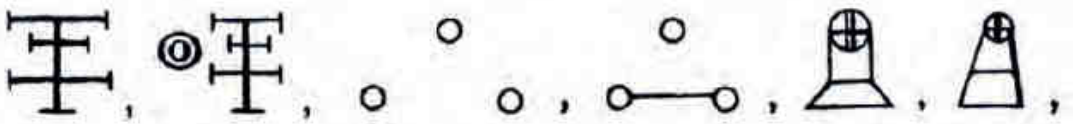
Although Gessmann did not cite any sources, there can be no doubt that he had found the symbols 4◦ and ♌ in Johann Christof Sommerhoff's *Lexicon Pharmaceutico-chymicum, Latino-germanicum Et Germanico-latinum* [Pharmaceutical-chemical Lexicon, Latin-German and German-Latin], which had been published in Nuremberg in 1701. Gessmann's and Sommerhoff's references to Sennert's oil of tartar and to urine, and their reproductions of the corresponding alchemical symbols, are reproduced in figure 3.5.

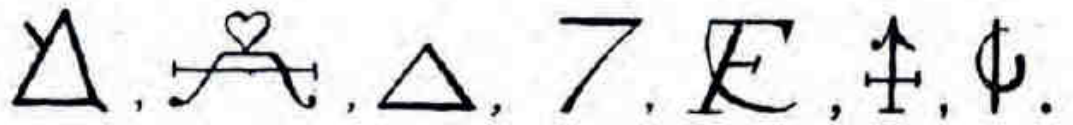
In both Gessmann's and Sommerhoff's works, the references to Sennert can only be to Daniel Sennert or Danielis Sennerti Vratislaviensis (1572–1637), Doctor and Professor of Medicine at the Academy of Wittenberg in Saxony. Sennert was well known as an experimental scientist, especially for

his work on tartar salt (potassium carbonate) as a medicine. We have not found any reproduction of a symbol for oil of tartar in any of Sennert's works. But if he used the symbol , this places the origins of the symbol no later than 1637. Whether it was used by other, or earlier, scientists than Sennert, we do not know.

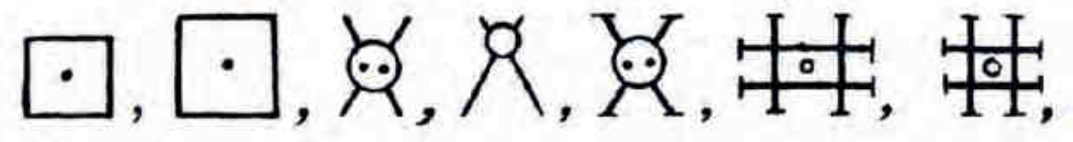
Weinsteinöl, *Sennerts; lat.: Oleum tartari Sennerti; fr.: Huile de tartre S.; e.: Tartar-oil of S.; it.: Olio tartaro di S.*

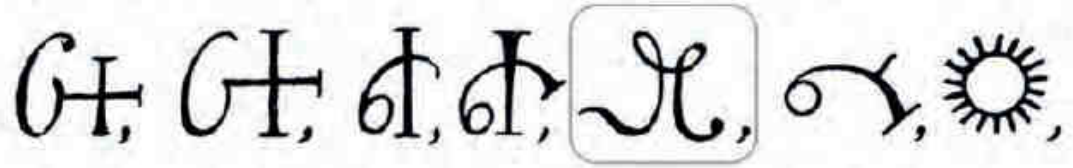


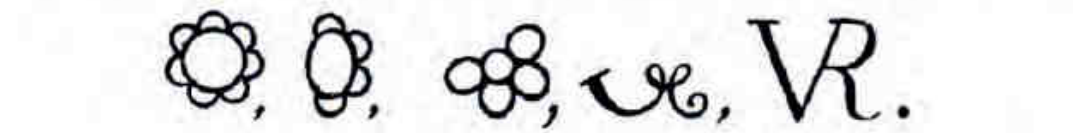


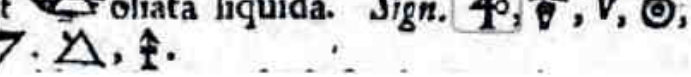


Harn, *Urin; lat.: Urina. lotium; fr.: Urine; e.: Urin; it.: Urino.*







Oleum Tartari Sennerti, eſt oliata liquida. *Sign.* 

Urina ⊡. Germ. Urin, eſt Sal reſolutum generatum in Hepate (in quo fallun-tur) per ſuum Emunctorium, ut ſuperfluum excrementum Salis, à Natura pulſum. *Dorn. M. Rul. & Johnſ.* Sign.

Figure 3.5. Alchemical symbols for Sennert's oil of tartar and urine. *Above: from Gessmann, G. W. (1922), pages LXXV and XXV; below: from Sommerhoff, Johann Christof (1701), pages 262 and 403; additional graphics by author*

None of this means that alchemical terms are pervasive in the Voynich manuscript. In fact, our interpretation is the reverse. The Voynich manuscript is not an alchemical text. More probably, the authors were familiar with alchemical symbols, and they used these particular symbols among others in their instructions to the scribes for the construction of the text.

Increasingly, it seems to us that the Voynich scribes were working from a document or documents in a natural language, which they transliterated letter by letter, or syllable by syllable, using an invented alphabet. That alphabet included Latin letters, Arabic numerals, Latin abbreviations, alchemical symbols, and whatever other ingredients had taken the authors' fancy. Why they should do this, is part of the Voynich mystery.

THE RARE GLYPHS

In passing, the v101 alphabet also identifies eighty-six "rare glyphs," each of which occurs only once or a few times in the whole manuscript. For these glyphs, the v101 keyboard assignments are "high ASCII codes" represented by expressions such as @243;—in other words an expression starting with @, followed by three numerals and a semi-colon.

By this author's count, there only 341 occurrences of "rare glyphs" in the whole manuscript; within the scope of this book, they will not concern us.

With that, we may start the frequency analysis.

FREQUENCIES

Based on the v101 keyboard assignments, duly copied into Microsoft Word as unformatted text and with "rare glyphs" each replaced by a single

Unicode symbol, we can count the following:

- 32,879 spaces, each represented by a period (.)
- 2,460 “uncertain spaces,” each represented by a comma (,)
- 158,947 glyphs

We can now use the “find and replace” function in Microsoft Word to calculate the frequencies of all the v101 glyphs as well as the spaces. We have done this for all the glyphs represented in the basic v101 glyph set, but for simplicity in Table 3.5, we have shown the twenty most frequent glyphs, which account for 94.1 percent of all the glyphs in the entire manuscript. If we can find meaning in the text represented by these twenty glyphs, then surely the remaining 5.9 percent will follow.

We may now attempt the “Gold Bug” approach.

Fans of Edgar Allen Poe will know that “The Gold Bug,” first published in 1843, was probably Poe’s most successful short story. The plot is a standard treasure-hunting tale, but what made it innovative was that Poe, at the center of his narrative, placed a cryptogram.

The hero of the story, one William Legrand, receives an enciphered text consisted of seven lines of numerals and characters, of which the first three lines are:

53‡‡†305))6*;4826)4‡.)4‡);80
6*;48†8¶60))85;1‡(;:‡*8†83(88)
5*†;46(;88*96*?;8)*‡(;485);5*†.

Legrand cracks the cipher by a means which, in the modern era, is child’s play: the analysis of character frequencies. He uses the fact that, in the English language, the most-common letters, in descending order of frequency, are e, t, a, i, o, n, s, and so on. In any piece of enciphered English text (provided that the cipher is a simple substitution), there is a good chance that the most common character will represent an “e,” the second most common will be an “t,” and so on.

Legrand spots that in the ciphertext, the most common character is an "8," which must decode to an "e"; the second most common is ";" which he decodes to "t"; and so on. He deploys this approach to convert the character string into English letters, of which the first three lines (with spaces inserted at the appropriate places) eventually are revealed as:

> a good glass in the bishop's hostel
> in the devils seat forty one degrees
> and thirteen minutes northeast and …

The plain text is the key to the location of the treasure; and after a few false steps, Legrand duly finds the buried gold and jewels, worth millions of dollars.[20]

Sir Arthur Conan Doyle employed a virtually identical approach in the Sherlock Holmes story, "The Adventure of the Dancing Men," first published in 1903. In this tale, Holmes receives a series of pieces of paper bearing mysterious drawings of matchstick men who seem to be dancing. The first of these drawings is this:

Table 3.5. Voynich manuscript: frequencies of the twenty most common glyphs as defined by the v101 alphabet, and of spaces and uncertain spaces.

glyph	v101	count	percent	percent
glyphs and spaces		194,286	100.0%	-
space	.	32,879	16.9%	-
uncertain space	,	2,460	1.3%	-
glyph total		158,947	81.8%	100.0%
o	o	25,043	12.9%	15.8%
of which †o	4o	5,265	2.7%	3.3%
o other	o, other	19,778	10.2%	12.4%
9	9	17,436	9.0%	11.0%
a	a	14,496	7.5%	9.1%
c	c	13,750	7.1%	8.7%
ct	1	11,052	5.7%	7.0%

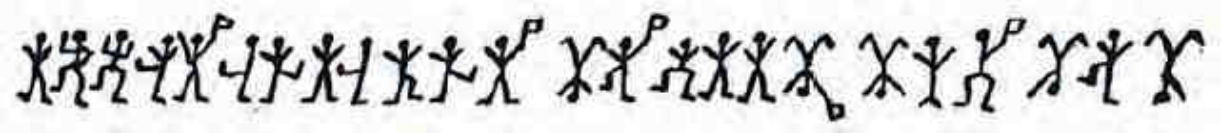
٤	e	10,652	5.5%	6.7%
٨	8	10,343	5.3%	6.5%
٣	h	9,919	5.1%	6.2%
٢	y	6,717	3.5%	4.2%
٣	k	6,118	3.1%	3.8%
٦	4	5,408	2.8%	3.4%
of which ٦٥	4o	5,265	2.7%	3.3%
٦ other	4, other	143	0.1%	0.1%
١١٦	m	4,112	2.1%	2.6%
٥٢	2	3,359	1.7%	2.1%
٥٥	C	2,844	1.5%	1.8%
٨	7	2,714	1.4%	1.7%
٢	s	2,702	1.4%	1.7%
١٦	n	1,770	0.9%	1.1%
٩	p	989	0.5%	0.6%
٥٣٣٥	K	910	0.5%	0.6%
٣٩	g	875	0.5%	0.6%

Source: text from Claston, Glen. *IVTFF v101 1.7, voyn101.txt*; analysis by author.
Note: ٦٥ {4o} is probably a single glyph.

Applying the frequency distribution of the most-common letters in English, Holmes sees that each of the fourth, sixth, ninth, and thirteenth of the dancing men probably represents an “e.” He then makes short work of decoding this message as

“Am here Abe Slaney”

The final message is contained in the following drawing.



Decoded by Holmes, it reads “Elsie prepare to meet thy God.” In due course, as per the Holmesian formula, a dastardly murder is solved, and a poor woman is saved.[21]

The variant of the “Gold Bug” approach that we need to apply to the Voynich manuscript is to compare the frequencies of the most-common Voynich characters (for our purposes, the top twenty or so should do) with the frequencies of the most-common letters in known languages. If we find a language whose frequency distribution approximates to that of the Voynich, maybe we have found one of the underlying languages of the Voynich. Or maybe not, but we shall see.

In the course of this approach, we shall have to consider whether, in the known languages, we separate out the letters with diacritics (accents) or not: for example, the “á,” “à,” “â,” and “ã” in French and Portuguese. Since the v101 basic glyph set has more glyphs than we expect, we are inclined to treat letters with diacritics as separate, for the time being.

As for the candidate languages, the distance-decay hypothesis gives us reason to think that medieval Italian and Latin should be among the top candidates. After these, we should look at medieval French, German, Bohemian (the precursor of modern Czech), and Greek, followed probably by medieval Iberian languages such as Spanish and Galician-Portuguese.

Here the issue arises, for any given language, as to what corpus of text we should use as the basis for calculating letter frequencies. Modern languages present no problem, since corpora of billions of characters are freely available on the internet, extracted from digitized newspapers, literature, and correspondence. Medieval languages are not so rich in computer-readable corpora. Appendix 4 lists a few which this author has found.

These corpora are designed to facilitate searches for words, and are not always well suited to search for individual letters. However, the search functions can usually be adapted for our purpose. For example, in the OVI corpus of medieval Italian, the search string *a* will yield all occurrences of the letter “a,” in upper- or lowercase.

Table 3.6. Frequencies of the twenty most common glyphs in the Voynich manuscript, and of the most common letters in medieval Italian

glyph	v101	Voynich manuscript	
		count	percent
total		158,947	100.0%
o	o	25,043	15.8%
9	9	17,436	11.0%
a	a	14,496	9.1%
c	c	13,750	8.7%
1	1	11,052	7.0%
e	e	10,652	6.7%
8	8	10,343	6.5%
h	h	9,919	6.2%
y	y	6,717	4.2%
k	k	6,118	3.8%
4	4	5,408	3.4%
m	m	4,112	2.6%
2	2	3,359	2.1%
C	C	2,844	1.8%
7	7	2,714	1.7%
s	s	2,702	1.7%
n	n	1,770	1.1%
p	p	989	0.6%
K	K	910	0.6%
g	g	875	0.6%

Latin letter	Italian, medieval (OVI)			
	OVI corpus		*La Divina Commedia*	
	count	percent	count	percent
all	110,622,089	100.0%	416,477	100.0%
E	13,949,820	12.6%	48,496	11.6%

A	11,048,819	10.0%	43,594	10.5%
I	10,385,126	9.4%	40,883	9.8%
O	10,307,988	9.3%	39,159	9.4%
N	7,476,593	6.8%	27,660	6.6%
R	7,221,090	6.5%	27,207	6.5%
L	6,806,137	6.2%	23,961	5.8%
T	5,974,825	5.4%	24,392	5.9%
S	5,487,626	5.0%	23,395	5.6%
D	5,276,970	4.8%	15,558	3.7%
C	5,264,580	4.8%	21,509	5.2%
U	4,072,799	3.7%	13,978	3.4%
M	3,254,365	2.9%	12,439	3.0%
P	3,213,188	2.9%	11,759	2.8%
V	2,038,020	1.8%	8,324	2.0%
G	2,038,010	1.8%	7,759	1.9%
H	1,701,520	1.5%	7,617	1.8%
F	1,433,382	1.3%	5,533	1.3%
Q	1,027,788	0.9%	3,364	0.8%
B	920,972	0.8%	3,241	0.8%

Source: Voynich v101 transliteration from Zandenberg, Rene. *IVTFF v101 1.7*; OVI text from Istituto Opera del Vocabolario Italiano, *Corpus OVI dell'Italiano antico*; *La Divina Commedia* from Project Gutenberg; analysis by author.

Without further ado, we may now attempt to match the Voynich glyph frequencies with the letter frequencies in medieval Italian, for which we have used two sources: the OVI corpus, and the full text of Dante Alighieri's *La Divina Commedia*. The results are shown in Table 3.6 and figure 3.6.

At first glance, these results look encouraging. The frequencies of the twenty most common Voynich glyphs seem to be quite close to those of the most-common letters in medieval Italian (and the Italian letter frequencies are very similar whether we use the large OVI corpus or the shorter text of *La Divina Commedia*.)

One anomaly stands out. The Voynich glyph ○ {o}, has a frequency of 15.6 percent, compared with 12.6 percent for the letter "e" in medieval Italian. Even if we include the variants of the Italian letter "e" that bear diacritics ("é" and "è"), we will still reach only 13.1 percent. The difference is rather large; and as we have seen, there are few European languages in which the most common letter has a frequency of over 15 percent.

Be that as it may, we have nothing to lose by applying the "Gold Bug" approach.

As a test, we tried this on the first three lines of folio f001r. The results are shown in Table 3.7.

We would not expect to get lucky on the first shot. The transliterations do not look much like Italian. Nevertheless, they do not look entirely like gibberish.

There are several ways to proceed:

- on the basis that the frequency rankings are not exact: try swapping letters with similar frequencies (for example, M with P, V with G)
- try another corpus of medieval Italian
- try another medieval language (Italian is clearly not the only candidate)
- and if none of this works, try modifying and streamlining the v101 keyboard assignments, along the lines that we suggested above (for example, combining some glyphs and disaggregating others).

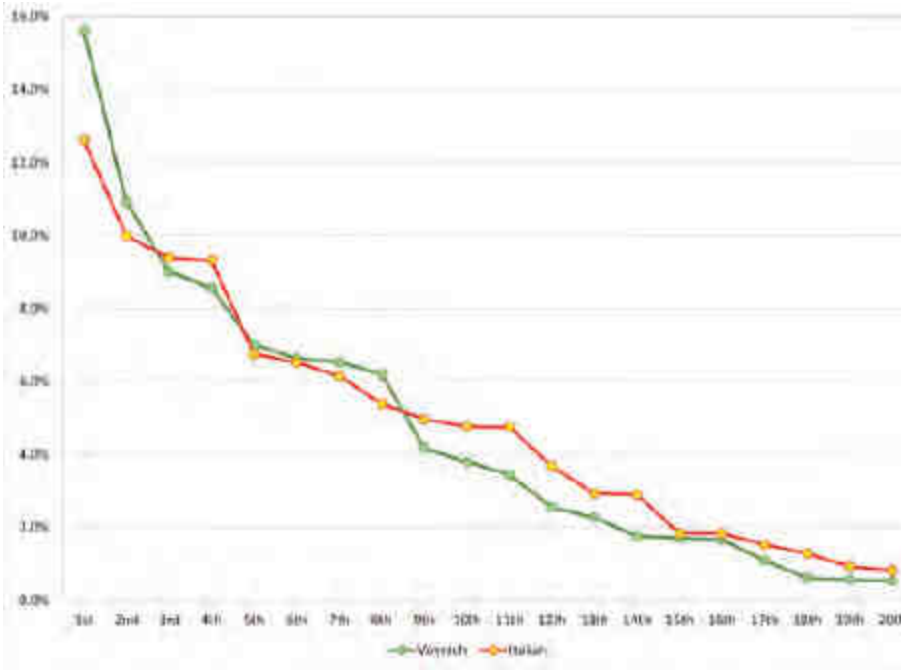


Figure 3.6. Voynich manuscript, frequencies of the twenty most common glyphs; and frequencies of the twenty most common letters in medieval Italian. *Source: author's analysis*

Our attempts to refine the "Gold Bug" process (no pun intended) will be documented in later chapters.

Incidentally (and for what it is worth), we may observe that on page f001r, the third line seems to rhyme with the first line (just as every third line does in every verse of *La Divina Commedia*).

Table 3.7. The Voynich manuscript, v101 alphabet: a test of the "Gold Bug" approach

page f001r, first three lines

Voynich v101 Unicode glyphs

fachedy9 fhay ar alfaiw chody chody 🌀ydy9 fhor chody89

dody9 dfhear or 9 fhacy cfhaiw chay ary dfhay dfhay 8ay

dgaiw? chcfh9 or 9fhaiw ch08 dfho ary9 dfhcy 8ay aiw dg

keyboard assignments as per v101 transliteration

fa19s.9,hae.ay.Akam.2oe.!oy9.@253;scs.9.hoy.2oe89

soy9.Hay.oy,9.hacy.1kam.2ay.Ais.Kay.Kay.8aN

s9aly.2ch9.oy.9ham.+o8.Ko,ay9.Kcs.8ay,am.s9

mapping to medieval Italian based on the OVI corpus

*INAM A LIR IS *CID PER *ESA 🌀MOM A LES PERTA

MESA FIS ES A LIOS NCID PIS **M VIS VIS TI*

MA!*S POLA ES ALID *ET VE ISA VOM TIS ID MA

mapping to medieval Italian based on *La Divina Commedia*

*INAM A TIR IS *CID PER *ESA 🌀MOM A TES PERLA

MESA FIS ES A TIOS NCID PIS **M VIS VIS LI*

MA!*S POTA ES ATID *EL VE ISA VOM LIS ID MA

Sources: author's analysis based on glyph and letter frequencies, with vowel-glyphs identified by Sukhotin algorithm (see Appendix 5). Image: *Beinecke Rare Book and Manuscript Library* . Notes: @253 (in v101) and 🌀 (in transliteration) denote a "rare glyph". The most frequent 20 glyphs were mapped; * denotes less frequent glyphs which were not mapped.

CHAPTER 4

ZIPF

Figure 4.1. George Kingsley Zipf. *Image by courtesy of Glottometrics, Fabian Bachrach*

Before we embark on our own voyage of investigation of the Voynich manuscript, we must consider the alarming possibility that it is a journey to nowhere. That is to say: it may be that the manuscript cannot be translated or deciphered because it has no intrinsic meaning. For want of better words, we must consider that the manuscript could be a hoax or a forgery.

There is plenty of precedent for such a situation. Throughout history (or so we imagine), fraudsters have created documents that purported to be what they were not: either artifacts of great value in their own right, or keys to vast treasures for those willing to pay the price.

One of the most celebrated examples of such fabrications is the Beale Cipher.

In the year 1885, the Virginia Book and Job Print company of Lynchburg, Virginia, on behalf of one J. B. Ward (who did not claim to be the author), offered for sale a twenty-three-page pamphlet. The title was “The Beale Papers, containing authentic statements regarding the Treasure

Buried in 1819 And 1821, near Bufords, in Bedford County, Virginia, and Which Has Never Been Recovered."[1]

The price of the pamphlet was fifty cents. If we believe the official inflation rates over the last century or more, fifty cents in 1885 would be equivalent to $13.56 in 2021. Put another way: in the mid-1880s in Ohio (and we suppose Virginia would not be much different), fifty cents would buy five pounds of granulated sugar, ten pounds of common soap, or forty pounds of corn meal.[2]

The unnamed author claimed to have been an associate of the late Robert Morriss, of Loudoun County, Virginia. The pamphlet contained a brief biography of Morriss, and the text of a series of letters dated between January and May 1822. The letters had been addressed to Morriss and signed "T.J.B." (the pamphlet giving it to be understood that these initials referred to a Thomas J. Beale, said to have been described by Morriss as "the handsomest man I had ever seen"). The author claimed to have received the letters as a bequest from Morriss, accompanied by a mandate to decipher their hidden message.

The hidden message in question took the form of three blocks of numbers, marked "1," "2," and "3," each block containing several hundred numbers, and each number having between one and four digits.

The first four lines of the block marked "2" are shown in figure 4.2.

In respect of the block marked "2," the author asserted that, after many failed attempts, he had found a tool of decipherment, in the form of the American Declaration of Independence, whereby he had decoded the numbers as follows:

> I have deposited in the county of Bedford, about four miles from Buford's, in an excavation or vault, six feet below the surface of the ground, the following articles, belonging jointly to the parties whose names are given in number three, herewith:
>
> The first deposit consisted of ten hundred and fourteen pounds of gold, and thirty-eight hundred and twelve pounds of silver, deposited Nov. eighteen nineteen. The second was made Dec. eighteen twenty-one and consisted of nineteen hundred

and seven pounds of gold, and twelve hundred and eighty-eight of silver; also jewels, obtained in St. Louis in exchange to save transportation, and valued at thirteen thousand dollars.

The above is securely packed in iron pots, with iron covers. The vault is roughly lined with stone, and the vessels rest on solid stone, and are covered with others. Paper number one describes the exact locality of the vault, so that no difficulty will be had in finding it.

The block of numbers marked "3" was labeled "Names and Residences," and supposedly identified the associates with whose help Beale had buried the fortune, and who were entitled to their respective shares in it.

The numbers marked "1" were labeled "The Locality of the Vault." The author claimed that he had devoted years of fruitless effort to decoding these numbers, at great cost to his health; and had eventually entrusted the task to an "agent," who we may suppose was the aforementioned J. B. Ward.

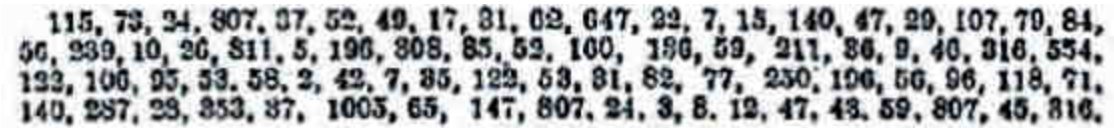

115, 73, 24, 807, 37, 52, 49, 17, 31, 62, 647, 22, 7, 15, 140, 47, 29, 107, 79, 84,
56, 239, 10, 26, 811, 5, 196, 308, 85, 52, 160, 136, 59, 211, 36, 9, 46, 316, 554,
122, 106, 95, 53, 58, 2, 42, 7, 35, 122, 53, 31, 82, 77, 250, 196, 56, 96, 118, 71,
140, 287, 28, 353, 37, 1005, 65, 147, 807, 24, 3, 8, 12, 47, 43, 59, 807, 45, 316,

Figure 4.2. The Beale Papers: the first four lines of block 2 of the enciphered numerals. *Public domain*

Needless to say, neither the Declaration of Independence, nor any variation of it, nor any other document, has served as a means of decoding block "1." Where Buford's Tavern used to stand, there is now only a ruined chimney; but over the years, within a radius of 4 miles in all directions, the county of Bedford has been invaded repeatedly by treasure hunters armed with dowsing wands, metal detectors and even earthmoving equipment, without the slightest success. May we be forgiven for taking the view that Beale's treasure does not, and never did, exist.

To return to the Voynich manuscript: if we are to devote our time and sweat to deciphering it, we need to have a reasonable conviction that it has an underlying meaning and is not a random collection of gibberish.

ZIPF'S LAW

For most of the century and more since Wilfrid Voynich's acquisition of the manuscript, the conventional wisdom has been that the document was indeed a text in an unidentified language, using an unknown script, and merely (if that is the right word) needed transliteration and translation. The basis for this belief was a linguistic rule of thumb known as Zipf's Law.

This rule was expounded by the American linguist George Kingsley Zipf in two academic monographs dated 1929[3] and 1932.[4] He called the rule, apparently interchangeably, the Theory of Relative Frequency or the Principle of Relative Frequency. It was subsequently named after him. Zipf wrote in engaging and accessible English; it is remarkable that his monographs, although they are available online, have rarely been quoted in his own words. It may be instructive, therefore, to read a few extracts from what Zipf wrote in 1929:

> Scholars long ago observed that whenever a word, hitherto uncommon, suddenly comes to be frequently used, it is likely to undergo a weakening of form to make it more easily pronounceable.
>
> … the weaker form is invariably the more frequent form.
>
> … Principle of Frequency. The accent, or degree of conspicuousness, of any word, syllable, or sound, is inversely proportionate to the relative frequency of that word, syllable, or sound, among its fellow words, syllables, or sounds, in the stream of spoken language.[5]

And, in 1932:

> The gist of the theory [of relative frequency] is that any element of speech which occurs in spoken language more frequently than some other similar element demands less emphasis or conspicuousness than that other element which occurs more rarely.

Zipf's idea of "emphasis" or "conspicuousness" meant the extent to which the word, syllable or sound mattered in the speech. Conversely, a "weak" word was one that did not matter. For example, in English the word "the" is extremely common but also very inconspicuous; we do not notice it, and many people leave it out or shorten it to the point of inaudibility. Yorkshire people (of whom this author is one) will know about phrases like "trouble down pit" or "going to t'pub." The word "accident" is rare but if used, is conspicuous: it gets your attention.

Zipf expressed his theory in mathematical terms as follows:

> Using X for frequency, and Y for conspicuousness, we can express our thesis thus:
>
> $Y = n/X$ or $XY = n$,
>
> where n is some constant.

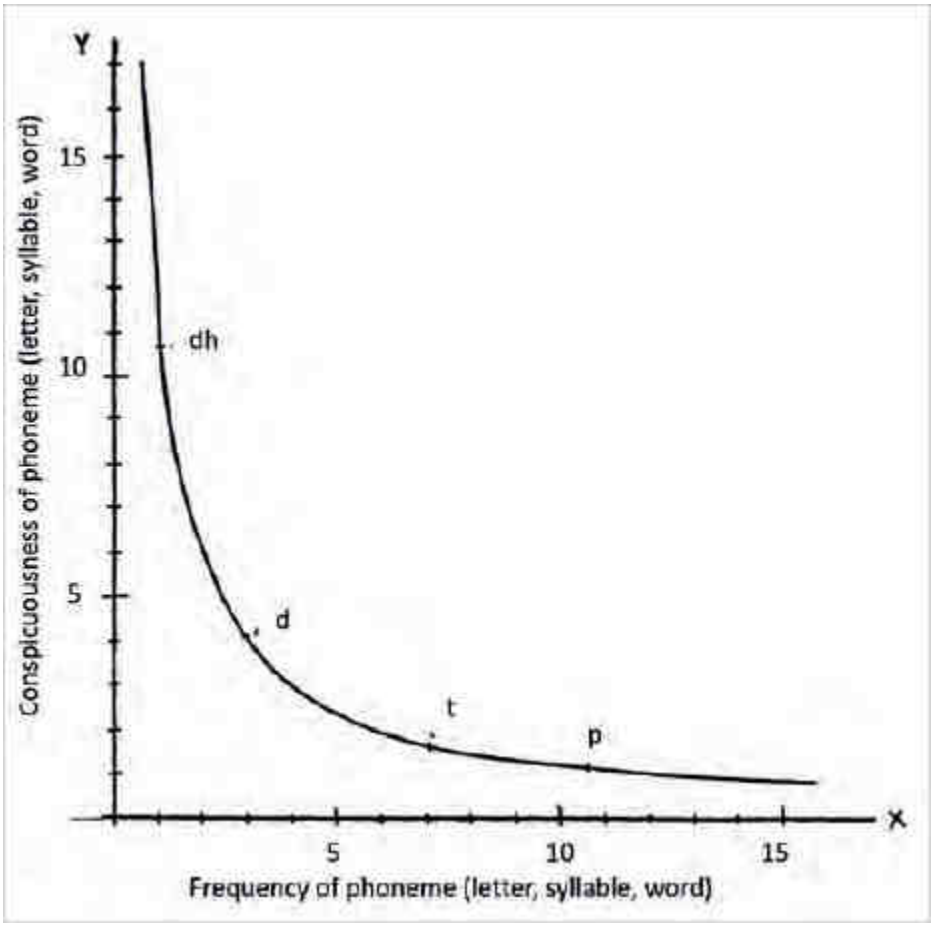


Figure 4.3. A graphical illustration of George Zipf's Principle of Relative Frequency, adapted from Zipf, George (1929), p.89, with additional graphics by author. *George Zipf / Harvard University Press*

To make it clear in visual terms, he presented the graph which is reproduced in figure 4.3. Mathematically minded readers will recognize the equation and the graph as representations of a hyperbola. Occasionally, hyperbolic equations enter our lives. Once in a while, an interstellar object (like Oumuamua) enters our solar system on a hyperbolic path, and leaves, never to be seen again.

Although Zipf put numbers on the Y axis of his graph, it is not obvious how he proposed to measure conspicuousness. At some point, his concept of "conspicuousness" mutated into "rank." It is not clear whether Zipf himself introduced this mutation (this author did not find any instance of the word "rank" in his monographs of 1929 and 1932).

In any event, Zipf's Law, as popularly formulated today, states that in any natural human language, the frequency of a word is inversely proportional to its ranking by frequency. For example, whatever is the most common word, the next three most common words should have the following frequencies:

second most common word	50.0% (½) of frequency of first word
third most common word	33.3% (⅓) of frequency of first word
fourth most common word	25.0% (¼) of frequency of first word

As an example, let us select the Brown University corpus of the American English language, and count the four most common words, as follows:

most common word	"the"	69,971 times

second most common word	“of”	36,412 times	52.0% of frequency of first word
third most common word	“and”	28,870 times	41.3% of frequency of first word
fourth most common word	“a”	23,225 times[6]	33.1% of frequency of first word.

As we can see, the proportionality is not very precise: “and” should be closer to 33 percent, and “a” should be closer to 25 percent. But we get the idea.

In order to test this rule on the Voynich manuscript, we need first to decide what constitutes a “word” in Voynich. As we explained in chapter 3, this is not self-evident. Certainly, the presumed text of the manuscript consists of strings of glyphs delimited by spaces, or by line breaks. But analysts have argued as to whether the strings are words, or syllables, or sounds; and even whether the spaces have some meaning. Can a word wrap from one line to the next? And if so, does it need something like a hyphen? We do not know. For simplicity, let us assume that a string surrounded by spaces (including “uncertain spaces”), or by a space and a line break, is a “word.” Then the twenty most common “words” in the Voynich manuscript are as shown in Table 4.1.

This sequence does not conform to Zipf’s Law at all: far from it. The word counts and frequencies are nothing like the sequence 100 percent – 50 percent – 33 percent – 25 percent, and so on, that Zipf’s Law predicts. To cite the most egregious example: the two most common “words,” ɑ ıı◌ {am} and ȣɑ ıı◌ {8am} have exactly the same frequency; by Zipf’s Law, one should be half as common as the other. The third most common “word,” oɋ {oe}, should be one-third as frequent; instead, its frequency is 96.1 percent of that of the first two.

Table 4.1. The twenty most frequent “words” in the Voynich manuscript

rank	“word”	v101	count	as % of 1st “word”
1	ȣɑ ıı◌	8am	739	100.0%
1	ɑ ıı◌	am	739	100.0%

3	oɋ	oe	710	96.1%
4	ɑƨ	ay	489	66.2%
5	oƨ	oy	440	59.5%
6	ƨ	s	432	58.5%
7	ɔɔoɋ	1oe	380	51.4%
8	ɔɔc9	1c9	375	50.7%
9	ɑɋ	ae	361	48.8%
10	ɔɔc89	1c89	360	48.7%
11	9	9	315	42.6%
12	89	89	294	39.8%
13	8ɑƨ	8ay	290	39.2%
14	ƨ	y	278	37.6%
15	ɋ	e	270	36.5%
16	4oIIɑıJ	4ohan	264	35.7%
17	4oIIɑıı	4oham	233	31.5%
18	4oIIcc9	4ohC9	232	31.4%
19	8ɑɋ	8ae	222	30.0%
20	ɔɔoƨ	1oy	217	29.4%

Source: author's analysis based on v101 transliteration. "Words" are defined as glyph strings delimited by spaces, uncertain spaces or line breaks. "Word" counts are by author from v101 transliteration in Microsoft Word and are case-sensitive by reference to v101.

We were therefore impelled to compare the Voynich "word" frequencies with those in other medieval documents, to see whether the Voynich manuscript was unusual in its divergence from Zipf's Law. We selected the Galician-Portuguese *Cantigas d'Amigo* (thirteenth century); Dante's *La Divina Commedia* (1320); Bocaccio's *Il Decameron, Prima Giornata* (1353); and John Wycliffe's *Holy Bible, New Testament* (ca. 1382–1395). The results are shown in figure 4.4.

From this graph we can see that, of the documents that we selected, the Voynich manuscript is by a long way the most divergent from Zipf's Law. The most conformal to Zipf's Law is Wycliffe's *New Testament*. The

document that, in this respect, most resembles the Voynich manuscript is *Cantigas d'Amigo*: a work of poetry.

Unlike English and most phonetic languages where the most-common words are short, several of the most common Voynich "words" are relatively long: for example, the five-glyph strings ꟼoⁿaiJ {4ohan}, ꟼoⁿccg {4ohC9} and ꟼoⁿaiiJ {4oham}. It is rather as if the words "would," "there," and "their" were among the top twenty most common words in English (whereas in the *Oxford English Corpus*, they are ranked respectively thirty-seventh, thirty-eighth, and thirty-ninth).

In fact, some of the most common Voynich "words" might be even longer than they appear in v101. We observed in chapter 3 that some v101 assignments could be interpreted as multiple glyphs. Perhaps iiJ {m} is not one glyph but three; in which case the "word" aiiJ {am} has four glyphs, and 8aiiJ {8am} has five. That would make the Voynich manuscript even more unusual in what Zipf would have called its "inconspicuous" words.

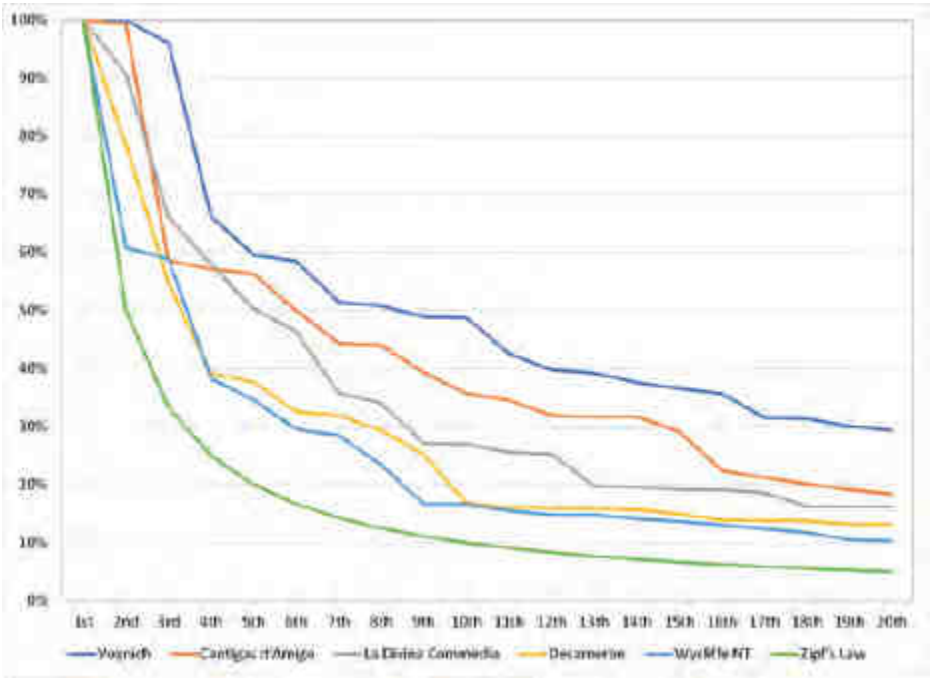


Figure 4.4. Voynich manuscript, frequencies of the twenty most common "words," compared with word frequencies in selected comparable documents, and the predictions of Zipf's Law. *Source: author's analysis*

Conversely, we have already considered that ᛏᛟ {4o} might be one glyph and not two. That would shorten some of the long glyph strings, although not by much. Also, we might wonder whether long strings such as ᛏᛟᛁᚠᚨᚾᛋ {4ohan} could be two "words," habitually spliced together. That kind of splicing occurs in Arabic, where the definite article ال ("the") is invariably attached to the noun to which it refers; but the ᛏᛟ {4o} is practically always the leftmost part of the "word," while in Arabic ال is always the rightmost syllable.

Perhaps the most reasonable conclusion that we can draw, at this point, is that the Voynich manuscript, as a whole, is substantially divergent from Zipf's Law, but not much more than other medieval documents that we have examined, especially if we consider documents with a poetic structure.

GLYPHS AS WORDS …

This may be a good a place as any to consider another test of Zipf's Law. Up to now, we have been assuming that glyph strings, delimited by spaces or line breaks, are words as we understand them in natural languages. But there are natural languages in which each character is a word: notably Chinese. What if the Voynich glyphs are words?

This is an easy hypothesis to test. We have the counts and frequencies of all the Voynich glyphs, as defined by the v101 alphabet. Let us assume that each glyph represents a unique word. Logically we should assume that the space is also a word; otherwise, there would be no need for spaces. If so, the space is the most common word, and the frequencies of all the glyphs should be measured against the frequency of the space. Then we can graph the frequencies of the space and the most-common glyphs against Zipf's Law. The results are shown in Figure 4.5.

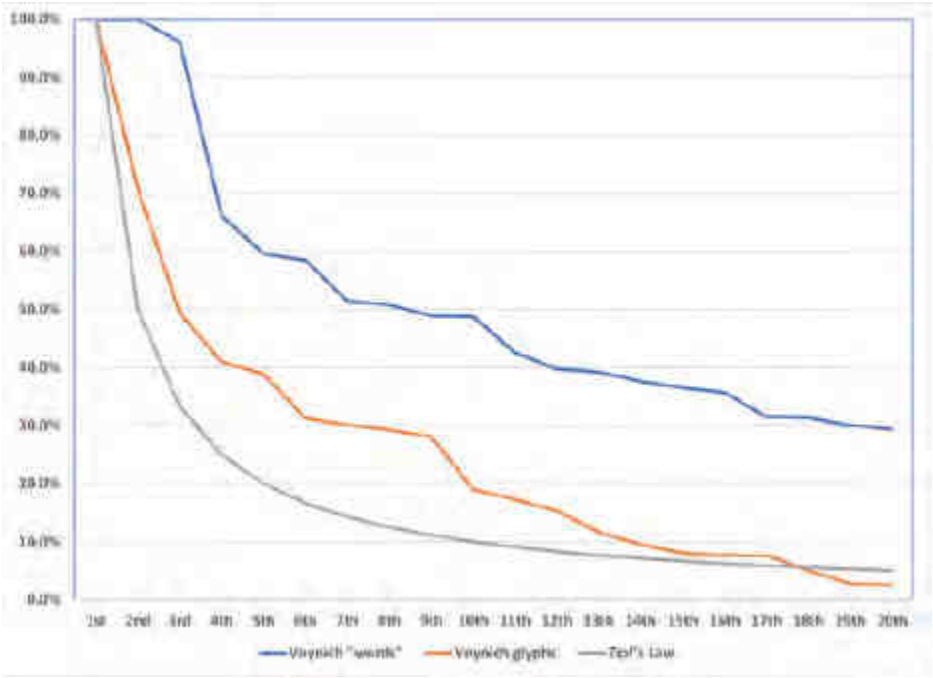


Figure 4.5. Voynich manuscript, frequencies of the twenty most common glyphs, compared with the frequencies of the twenty most common "words," and the predictions of Zipf's Law. *Source: author's analysis*

Extraordinarily, the Voynich glyphs conform vastly better to Zipf's Law than the Voynich "words."

It does not follow that the Voynich manuscript is written in pictograms, like Chinese. The Voynich glyphs conform better to Zipf's Law than the words of *Cantigas D'Amigo* or *La Divina Commedia*, but not as well as the words of *Decameron* or *The Gospel According to Matthew*. We can see from figure 4.5 that the frequencies of the third through ninth most common glyphs are quite divergent from the predictions of Zipf's Law.

The greatest objection to the pictogram hypothesis is that there are, at most, about 250 distinct glyphs in the Voynich manuscript, even if we include all the "rare glyphs" identified in Claston's v101 alphabet. By comparison, modern conversational Chinese needs a vocabulary of at least 2,000 words. A document with a vocabulary of about 250 words would be little more than a child's picture book.

In any case, there is a simple test of the pictogram hypothesis. If the Voynich space and glyphs are the words of some underlying language, their frequencies should approximate to those of the words in that language. We could try medieval Italian, in which the word frequencies are well

documented. In that case, the twenty most common Voynich glyphs and the most-common Italian words should match up as follows:

glyph: space o 9 a c 1 e 8 h y k 4 m 2 C 7 s n p K
v101: . o 9 a c 1 e 8 h y k 4 m 2 C 7 s n p K
Italian: e che di la a per in non è de lo et il si li le del se da con

We then apply this mapping to, for example, the first line of f001r, which starts with the v101 transliteration “fa19s.9,hae.ay.Akam.2oe.” The result is as follows:

{f} la per di del e di e è la in e la de e {A} lo la il e si che in

We can see that this makes no sense; nor can we imagine any permutation of the frequencies that might make sense. There are too many prepositions and conjunctions, not enough verbs, and no nouns. The top twenty words of any natural language would give a similar result; but in the Voynich manuscript, the space and the top twenty glyphs account for over 90 percent of the text. In short, the pictogram hypothesis does not work.

ON THE OTHER HAND …

For most of the modern history of the Voynich manuscript, it has been widely believed that the frequencies of “words” in the Voynich manuscript conformed to Zipf’s Law, and that therefore that the manuscript was based on natural languages.

The first robust challenge to this orthodoxy, as far as we could determine, came in 2004 from an academic in England: Professor Gordon Rugg of Keele University.

Rugg and his associate Gavin Taylor, of Heath Lane Academy in Leicester, later wrote:[7]

> [The possibility] that the manuscript contained only meaningless gibberish, had been generally discounted, because there are numerous odd statistical regularities in the text of the manuscript, which everyone believed were much too complex to have been hoaxed.
>
> Gordon [Rugg]'s work showed that this belief was mistaken, and that the most distinctive qualitative features of the Voynich Manuscript could be replicated using low-tech hoaxing methods.
>
> … Gordon's initial work, however, did not address the quantitative statistical regularities of the text in the manuscript. Our recent article in Cryptologia addresses this issue, and shows how the most distinctive quantitative features of the VMS can be replicated using the same low-tech hoaxing methods …

Rugg and Taylor's "low-tech hoaxing method" could be summarized as follows:

- to draw a tabular array of cells (like an Excel spreadsheet) on a flat surface
- to fill in some cells with Voynich glyphs and to leave others blank
- to construct a grille from a flat rigid board, with rectangular holes cut in various places (what is known as a "Cardan grille")
- to slide the grille over the array in a predetermined sequence, and record whatever Voynich glyphs, or blanks, showed up through the holes.

Figure 4.6. An example of a Cardan grille applied to the Voynich glyph set as shown in Table 3.1 This grille randomly selects up to seven glyphs from each of the rows. *Source: author's analysis*

Rugg and Taylor wrote:

> … we were able to show via Gavin [Taylor]'s work that three different sets of meaningless gibberish produced using the table and grille technique all follow Zipf's law, and are comparable to a range of real natural language texts in their Zipf's law curves.
>
> … we demonstrate[d] that the distribution of common syllables within the Voynich Manuscript is very different from the distribution of common syllables in real human languages, but very similar to what would be predicted from the production of meaningless gibberish produced using tables and grilles, with several changes of table for logistical reasons.
>
> … all the most striking features of the Voynich Manuscript can be produced by using tables and grilles to produce meaningless gibberish.[8]

Rugg and Taylor were not arguing that the Voynich was necessarily a hoax; they were arguing that it could have been a hoax. They were saying that the whole of the Voynich manuscript could have been generated by a mechanical and mechanistic process: what we might call an algorithm.

This author corresponded with Rugg and Taylor with a view to testing the viability of the "algorithm" hypothesis.

In the first instance, this author asked Rugg and Taylor whether their table and grille methodology would lend it itself to automation. It seemed that it would be feasible, and probably not difficult, to write a program in which the inputs were the Voynich glyphs (on some appropriate definition) and the output was a manuscript of the same length as the Voynich manuscript, with the same thematic sections. The program would need to replicate (but fuzzily, i.e., with some randomness) all the commonly known "laws" about, for example:

- the lengths of words
- the lengths of lines
- the differences between the thematic sections of the Voynich manuscript
- the differences between the Voynich A and B languages (of which, more later)
- and for each glyph, its frequency of occurring (a) at the beginning of a word (b) at the end of a word (c) in the middle of a word (d) at the beginning of a line (e) in the first half of the line (f) before or after or in conjunction with another glyph, and so on

It would then be instructive to make a statistical comparison of the program's output with the actual manuscript.

They replied:

> We found that the output was very sensitive to tweaks in the initial parameters, with apparently small changes in the table creation algorithm producing strikingly different outputs, even using the same initial set of syllables to populate the tables…. using the table and grille approach manually made sense of some odd features of Voynichese, as a result of human factors such as errors or shortcuts when tired, that would not show up in an automated version of the approach.[9]

Rugg and Taylor had demonstrated that an algorithm (namely a structured table of glyphs combined with a Cardan grille) could produce a manuscript which had many of the statistical attributes of the VMS; and that a simple algorithm could generate the appearance of complexity in the output.

This author was reminded of the Mandelbrot set: a mathematical shape which has been described as "the most complex object in the universe," but which is generated by a relatively simple algebraic algorithm.

It is still interesting to imagine an automated process to replicate the Voynich manuscript. For inspiration, one can turn to the short story "Pierre Menard, Author of the *Quixote*" by the Argentine author Jorge Luis Borges. Pierre Menard is a fictional French writer who, around the year 1934, sets out not to translate but to recreate Cervantes' *Don Quixote*; and ends by writing a *Quixote*, in seventeenth-century Spanish, which is identical to the original, but at the same time is, in a sense, original.

One can envisage the development of a computer program which would be run many times (at a guess, thousands of times), with variations on the parameters, and which would emulate a scenario that might have taken place sometime in the early fifteenth century (or possibly later, using a stock of fifteenth-century parchment).

In this scenario, there is a producer (possibly resembling J. B. Ward of the *Beale Papers*). He or she has probably several accomplices whom we might call "assembly workers" or simply "scribes."

The workplace preparation would have been along the following lines:

- Given the amount of parchment at his or her disposition, the producer sets out to create a manuscript of about 240 pages divided into seven sections.
- The producer intends to develop slightly different algorithms for each of the seven sections.
- The producer develops a glyph set or "alphabet" containing all the glyphs to be used. Possibly there are seven subsets with substantial overlap but with slight variations, corresponding to the seven sections.

- The producer constructs seven sets of dice, one for each of the seven sections of the manuscript. (Since dice have existed for thousands of years, the idea of dice is a simple alternative to a table and grille and avoids an assumption about the date when this process took place.)
- For each section, the corresponding set of dice contains enough individual dice for every glyph (needed for that section) to be inscribed at least once on a face of a die.
- Each die has six faces. (The program could possibly vary this assumption.)
- On each face there is inscribed (a) nothing (b) a single glyph (c) a pair of glyphs (d) a triplet or more glyphs, in some proportions determined by the producer.

The production process would have been approximately as follows:

- In preparation for creating a line, a scribe takes several dice from the set. This number is constrained in some way (for example, not more than four dice); the program must embody and vary the constraints.
- The scribe rolls the dice and records, in rough copy, the glyphs that appear on the top faces.
- There are several algorithms for dealing with blank faces:
    - treat a blank face as a space between "words"
    - ignore blank faces and concatenate the glyphs on nonblank faces.
    - if all faces are blank, roll the dice again
    - and so on …
- Conversely, to create spaces between "words," the scribe can either use a blank face from a roll of the dice or insert a space after every roll of the dice in which no blank faces have appeared.
- When sufficient glyphs and spaces have been recorded to complete a line, the scribe transcribes the line from the rough

copy to a good copy. This good copy is either on an intermediate document for approval by the quality controller, or on the final document; this issue is not critical but addresses the "functionality of the line" discovered by Prescott Currier (of which, more in Chapter 5).

- After a line has been completed, the scribe starts another series of rolls of the dice for the subsequent line.

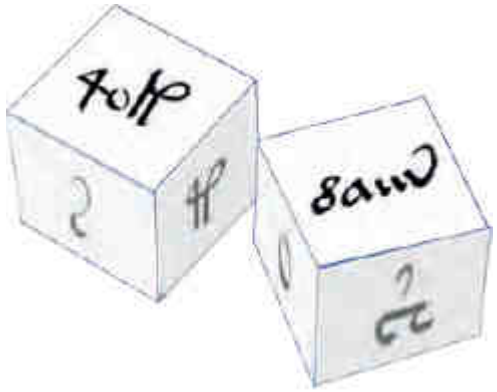


Figure 4.6A. An illustration of a hypothetical set of dice used to generate random glyph strings. *Source: graphic by author*

This process continues until enough text has been generated to fill the 227 pages, apart from spaces on which illustrations have already been drawn (perhaps even by a previous owner of the parchment).

The producer would have needed a strategy for writing on the faces of the dice. It might have been productive or even necessary to include not only single glyphs, but also pairs and triplets of glyphs, since there are clearly pairs and triplets that occur vastly more often than others.

It would not be essential for this process to be based on dice; but it is easy to visualize a process based on dice. We suspect that any randomizing process could be constructed from a combination of dice-based processes. As an illustration, for anyone who (like this author) does not know programming languages, Excel provides a simple randomizing process. For example:

- The formula =1+INT(6*RAND()) yields a random integer between 1 and 6.

- More generally, =1+INT(n*RAND()) yields a random integer between 1 and n (thus emulating the throw of an n-sided die).
- It would be straightforward to generalize the function to emulate multiple throws of dice.
- And one could further generalize to a randomized selection of a fixed or flexible number of dice from a larger set of dice (which would emulate grabbing a handful of dice from a bowl).

THE HOAX HYPOTHESIS

With this process in mind, one can advance a "hoax hypothesis" along the following lines:

- The Voynich manuscript was constructed by means of an algorithm or several algorithms (maybe up to seven, corresponding to the seven sections; or maybe two, corresponding to the two Voynich languages).
- We could think of each algorithm as having three components:
  - structure (meaning, to pursue the dice analogy, concepts such as the number of faces on each die; the number of bowls of dice; the number of dice drawn from the bowl for each throw)
  - content (meaning what is written on the faces of the dice)
  - randomization (meaning the throwing of the dice).
- The functional unit of the manuscript is the line (in most cases, a string of characters delimited by the width of the page).
- The line may have the function of quality control. That is, a scribe creates a line in rough copy, passes it for approval or amendment to a quality controller, who passes it to a scribe (the same or another) for writing the fair copy.
- We could conceive that there are no functional units smaller than the line; in other terms, that there are no "words" in the Voynich manuscript. The strings that resemble natural-language

> words (i.e., strings of glyphs preceded or followed by spaces) have no function. The spaces are generated by the algorithm (to return to the dice analogy: a blank face on a die generates a space in the string).

The structure of the algorithm consists of, or can be emulated by, the following:

- defining the number of faces per die (which could be six, or could be another number)
- creating a set (a "bowl") of dice which contains at a minimum all the individual glyphs, and very probably, all or most of the pairs of glyphs; and possibly some triplets or longer combinations of glyphs, in the frequencies in which they are intended to appear in the script; possibly, a few hundred dice are needed
- delimiting the number of dice that are to be drawn from the bowl for any one throw (a number such as four or five might be reasonable).

The randomizing process is simply the throwing of the dice.

In this hypothesis, there is no decision making, except when the end of a line is reached. Frequency or rarity of glyphs is handled by the content of the faces.

In correspondence with this author, Rugg and Taylor advised acquiring some hands-on experience of producing text manually using the materials available in the fifteenth and sixteenth centuries. They wrote:

> [We] considered dice, coins and prism-shaped wands (all used in mediaeval gaming) as possible mechanisms for producing the manuscript, but ended up concluding that tables and grilles were much less effort to use once the tables and grilles were set up.[10]

This author did search the local shops for blank rewritable dice but did not find any. Maybe that was just as well. The "algorithm hypothesis" is almost believable. It seems clear that an algorithm could have created something resembling the Voynich manuscript, in a statistical sense; and it is possible to imagine the process whereby this was done. But it could never be proved that this was the origin of the manuscript. Unlike Pierre Menard in Borges's *Author of the Quixote*, no algorithm could produce a manuscript identical with the Voynich. In any case, it is more interesting to assume that the manuscript has meaning, than that it has none. This author's search for the key to unlock that meaning will be narrated in the chapters that follow.

A TEST OF THE ALGORITHM

Nevertheless if, as Professor Rugg proposed, the Voynich manuscript could be re-created by means of an algorithm, we had to put this hypothesis to the test.

For this purpose, we did not need to use rewritable dice. The random number function in Excel was sufficient.

Our test of the algorithm was as follows:

- We set out to re-create the first page of the Voynich manuscript, page f001r.
- We took it that this page was written in Prescott Currier's Language A (of which, more in Chapter 5). According to the v101 transliteration, the pages of the Voynich manuscript that are written in Language A contain 55,548 characters, consisting of 45,499 glyphs plus 10,049 spaces and uncertain spaces. The frequencies of each glyph, and of the spaces and uncertain spaces, are known.
- We created a virtual "dice bowl," consisting of 528 characters (equivalent to approximately 1 percent of the Language A corpus). The "bowl" contained the thirty-two most frequent characters in Language A. For example:

  - the “space” character, which occurred 9,428 times in Language A, was present ninety-four times in the “bowl”
  - the glyph o {o}, which occurred 8,643 times, was present eighty-six times in the “bowl”
  - the glyph o {9}, which occurred 4,435 times, was present forty-four times in the “bowl”
  - and so on, down to glyphs which occurred between fifty and 100 times and were present only once in the “bowl.”
- The characters in the “bowl” were numbered from 1 to 528.
- In other words, the “bowl” was a representative sample of the corpus of Language A, missing only the “rare glyphs.”
- We set up an Excel spreadsheet with dimensions matching those of page f001r, and filled the cells with random integers.

By this means, we created a replica of page f001r, with the same number of lines, the same number of characters on each line (including glyphs and spaces), but with randomly generated characters where the original characters had been. There was no human intervention in the selection of glyphs or spaces. Both the original and the replica contained 968 characters including spaces.

In statistical terms, there was one important difference between the replica f001r and the original f001r, and one significant similarity.

The important difference was that the replica, being allowed to generate spaces randomly, occasionally generated two consecutive spaces, or a space followed by an uncertain space: and conversely, occasionally very long strings of glyphs. Consequently, in the replica, the average word length was longer: 5.12 glyphs, versus 3.72 glyphs in the original.

The similarity was that the frequencies of the glyphs in the replica were a good match for the frequencies in the original (as we expected, since our “dice bowl” was designed to be representative of the Language A corpus). The correlation between the frequencies in the original and the replica was 93.1 percent.

The statistical comparison between the original and the replica is summarized in figure 4.7.

This test illustrated that an algorithm based on random numbers could produce a document that statistically resembled the Voynich manuscript. To create a better statistical resemblance, the algorithm would have to incorporate some rules for keeping the lengths of words comparable with those in the original (which the simple algorithm above does not).

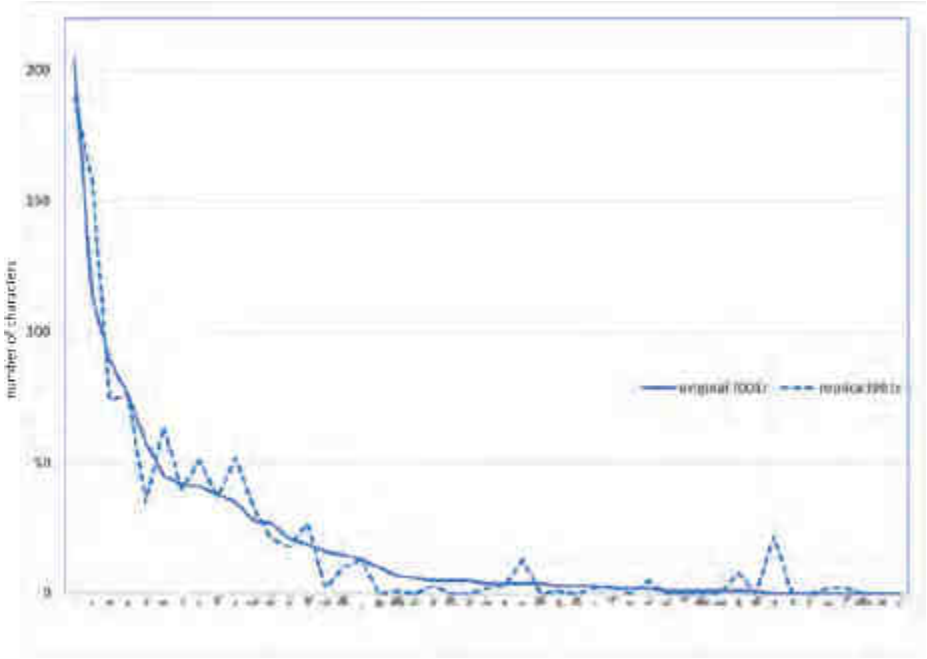


Figure 4.7. Statistical comparison between page f001r of the Voynich manuscript and a replica of page f001r created by an algorithm. *Source: author's analysis*

However, when we look at the strings that such an algorithm produces, we see that there is something missing. Figure 4.8 shows how the replica's first three lines compares with the original.

While the replica and the original have the same number of characters on each line (including glyphs and spaces), visually the replica does not greatly resemble the original. Most strikingly, it contains strings such as {1hk} and {mmm}, which rarely occur in the Voynich manuscript. Randomizing processes permit such strings, as any poker player knows who has ever held jack-queen-king, or three of a kind; but the rules of natural language generally do not. In most European languages, letters can be doubled but not tripled, and nearly all letters have other letters which they cannot follow or precede.

If the manuscript was created by an algorithm, it was one in which someone put considerable effort into incorporating the rules of natural language.

With that, we leave the idea of an algorithm aside, and continue our quest for a natural language or languages which might have been the source, or the inspiration, for the scribes of the Voynich manuscript.

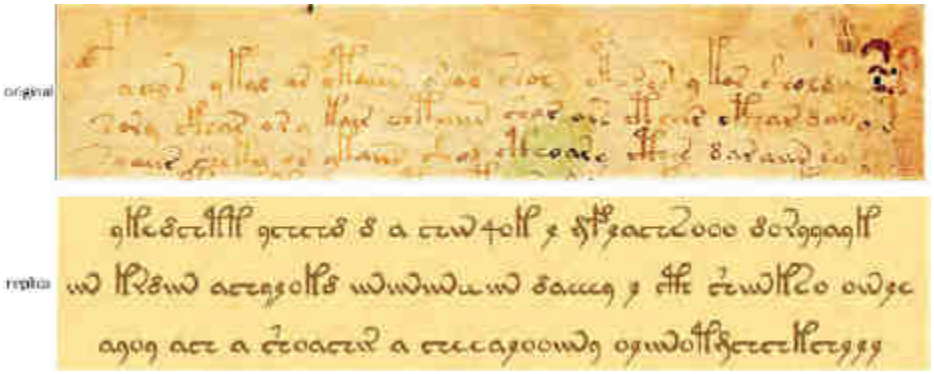


Figure 4.8. Voynich manuscript: the first three lines of page f001r, and a re-creation of these lines generated by an algorithm based on random numbers. *Beinecke Rare Book and Manuscript Library, additional graphics by author*

CHAPTER 5

CURRIER

Figure 5.1. Capt. Prescott H. Currier, ca. 1970s. *Station Hypo*

Captain Prescott H. Currier was one of the great unsung heroes of Voynich research. Single-handedly, he developed several substantive hypotheses on the Voynich manuscript that have withstood the test of time (and very few people could make that claim).

In 1976, Brig. John H. Tiltman, one of the pioneering researchers of the Voynich manuscript, wrote as follows:

> To my knowledge there have been three rather extensive analyses of the script of the manuscript, by Mr. [William F.] Friedman, by me, and by Captain Prescott Currier. Of these, I believe Captain Currier's to be far the most complete.[1]

Capt. Currier's hypotheses included the following:

- that, if the manuscript was derived from natural language, then it was based on at least two, and maybe more
- that the manuscript had been written by several individuals who, even when using a common language, had observably different styles of handwriting
- and that the line had some function (that is to say, a line was not, as it is in normal prose, merely a chunk of text that happened to have wrapped from the previous line).

Currier was born on June 4, 1912, in Holbrook, Massachusetts. It was the same year in which Wilfrid Voynich discovered the manuscript that he called the "Roger Bacon Cipher" in (as he would claim) a castle in southern Europe. In 1929, Currier enlisted in the US Navy. In 1932, he became a member of what was called the "On the Roof Gang," a group of radio operators who were trained in Washington, DC, in radio interception and cryptanalysis. He was sent to the Philippines where he learned Japanese. He completed four years of service and left the Navy to earn a bachelor's degree in Romance Languages[2] at George Washington University. In 1936, he returned to the Naval Reserve as a commissioned officer. He had a distinguished career in the Navy and as a civilian in the National Security Agency.[3]

On November 30, 1976, at the invitation of Mary D'Imperio, Capt. Currier was a speaker at a seminar in Washington, DC, titled "New Research on the Voynich Manuscript." The subject of his presentation was "Some Important New Statistical Findings."

He made the first two of his results clear at the outset:

> The two most important findings that I think I have made are the identification of more than one hand and the identification of more than one "language."
>
> … my use of the word language is convenient, but it does not have the same connotations as it would have in normal use.
>
> … There are two different series of agglomerations of symbols or letters, so that there are in fact two statistically distinguishable "languages."[4]

Currier called these two languages A and B. He went on to describe how he had found evidence of languages A and B in most of the thematic sections of the manuscript. He did not summarize the relationship between languages and sections; but he did contribute a handwritten appendix in which he recorded his identification of the language on most of the folios of the manuscript.

Our first observation, for what it is worth, is that every folio has a recto and a verso. Apparently, Currier's view was that every folio that he studied, without a single exception, was written in the same language on the recto and the verso.

Forty-four years would pass before Currier's identification of the languages was revisited. In 2020, Dr. Lisa Fagin Davis, of the Medieval Academy of America, made a meticulous reexamination of the Voynich manuscript, page by page. For nearly all of the pages for which Currier had made an identification of the language, Dr. Davis agreed with him. She identified just one folio in more than one language: on folio 57 she saw Language B on the recto and Language A on the verso.

In addition, Dr. Davis identified the language for all the other pages (and in nearly all such cases, she specified Language B). Currier's and Davis's identifications of the languages are summarized in Table 5.1.

From this table we can calculate the extent to which, in each of the thematic sections, Currier's languages A and B were used. Here we are referring to the themes as defined by René Zandbergen. It is worth recalling that the themes are not all in contiguous sections; also, that the manuscript pages appear not to be in their original order, and that some are missing. Our analysis is shown in Table 5.2.

Table 5.1. Voynich manuscript: Capt. Prescott Currier's and Dr Lisa Fagin Davis' analyses of the languages and hands

folio	Currier		Davis	
	language	hand	language	hand
001	A	1	A	1
002	A	1	A	1
003	A	1	A	1
004	A	1	A	1

005	A	1	A	1
006	A	1	A	1
007	A	1	A	1
008	A	1	A	1
009	A	1	A	1
010	A	1	A	1
011	A	1	A	1
012		missing		
013	A	1	A	1
014	A	1	A	1
015	A	1	A	1
016	A	1	A	1
017	A	1	A	1
018	A	1	A	1
019	A	1	A	1
020	A	1	A	1
021	A	1	A	1
022	A	1	A	1
023	A	1	A	1
024	A	1	A	1
025	A	1	A	1
026	B	2	B	2
027	A	1	A	1
028	A	1	A	1
029	A	1	A	1
030	A	1	A	1
031	B	2	B	2
032	A	1	A	1
033	B	2	B	1
034	B	2	B	2
035	A	1	A	1
036	A	1	A	1

037	A	1		A	1
038	A	1		A	1
039	B	2		B	2
040	B	2		B	2
041	B	2		B	5
042	A	1		A	1
043	B	2		B	2
044	A	1		A	1
045	A	1		A	1
046	B	2		B	2
047	A	1		A	1
048	B	2		B	5
049	A	1		A	1
050	B	2		B	2
051	A	1		A	1
052	A	1		A	1
053	A	1		A	1
054	A	1		A	1
055	B	2		B	2
056	A	1		B	1
057	B	2		rB, vA	r5, v1
058	A	-		B	3
059			missing		
060			missing		
061			missing		
062			missing		
063			missing		
064			missing		
065	-	-		B	3
066	B	-		B	5
067	-	-		B	4
068	-	-		B	4

069	-	-		B	4
070	-	-		B	4
071	-	-		B	4
072	-	-		B	4
073	-	-		B	4
074			missing		
075	B	2		B	2
076	B	2		B	2
077	B	2		B	2
078	B	2		B	2
079	B	2		B	2
080	B	2		B	2
081	B	2		B	2
082	B	2		B	2
083	B	2		B	2
084	B	2		B	2
085	B	3		B	4
086	B	3		B	4
087	A	4		A	1
088	A	4		A	1
089	A	-		A	1
090	A	-		A	1
091			missing		
092			missing		
093	A	4		A	1
094	B	5		B	3
095	B	5		B	3
096	A	4		A	1
097			missing		
098			missing		
099	A	4?		A	1
100	A	4?		A	1

101	A	4?		A	1
102	A	-		A	1
103	B	X		B	3
104	B	X		B	3
105	B	Y		B	3
106	B	X		B	3
107	B	-		B	3
108	B	-		B	3
109			missing		
110			missing		
111	B	-		B	3
112	B	-		B	3
113	B	-		B	3
114	B	-		B	3
115	B	-		B	2, 3[a]
116	rB, v-	-		B	3
Ros	-	-		A	4

Source: D'Imperio, Mary E. (ed.) (1976). *New Research on the Voynich Manuscript: Proceedings of a semina* r, 30 November 1976, Table A; Davis, Lisa Fagin (2020). "How Many Glyphs and How Many Scribes? Digital Paleography and the Voynich Manuscript," *Manuscript Studies* : Vol. 5 : Issue 1 (Spring 2020), Article 6. Note (a): first 12 lines in hand 2, rest in hand 3.

Table 5.2. The Currier languages used in each thematic section of the Voynich manuscript

thematic section	number of pages, by Currier language		
	total	Language A	Langua B
herbal	129	95	
astronomy	8	0	
cosmology	11	1	
astrology (zodiac)	12	0	
biology / balneology	19	0	

pharmaceutical	16	16
text with marginal stars	25	0
text only	7	1
total	227	113

Source: author's analysis based on Davis, Lisa Fagin (2020). "How Many Glyphs and How Many Scribes? Digital Paleography and the Voynich Manuscript" , *Manuscript Studies* : Vol. 5 : Issue 1 (Spring 2020), Article 6.

On the basis of Currier's identifications, Language A predominated in the "herbal" theme and was used for the whole of the "pharmaceutical" theme. Language B was used for about a quarter of the "herbal" theme, about half of the "cosmology" theme, the whole of the "biology" or "balneology" theme, and most of the text pages which had stars in the margins. In the "astronomy" and "zodiac" themes, Currier did not attempt to identify the language, but in both sections Dr. Davis identified the language as B.

THE DIFFERENCES BETWEEN A AND B

In his presentation to the Voynich seminar, Currier either did not wish, or did not have the time, to explain how he recognized the differences between languages A and B. The audience would surely have longed to know, but he did not say. What he divulged was that:

> … the characteristics of "languages" A and B are obviously statistical. (I can't show you what they are here, as I don't have slides prepared.)[5]

Surely, he must have counted the glyphs on each page that he studied, and he would have seen that certain glyphs were more frequent on some pages than on others. At some point he must have selected certain common

glyphs, and for each page that he studied, he decided of the following nature: if this set of glyphs occurs more frequently than usual, we are looking at language A; if this other set of glyphs is ubiquitous, we are confronted with language B.

Armed as we are today with computer-readable transliterations of the entire text, and with the search functions of Microsoft Word and Excel, we have reconstructed the differences that Currier must have observed. The process was as follows. There are several machine-readable versions of the entire text of the Voynich manuscript, mapped arbitrarily into American keyboard characters. For convenience, we used Glen Claston's mapping into the v101 alphabet. There was no need to count all the glyphs; it was sufficient to look at the following:

- the eighteen most common glyphs, which account for 94.0 percent of the text of the manuscript
- the "space" and "uncertain space," which in v101 are {.} and {,} respectively, in case they might also be glyphs hiding in plain sight.

It was then a matter of importing the v101 text into Microsoft Word; converting the ASCII strings for "rare glyphs" into a single Unicode symbol; splitting the text into three chunks corresponding to Language A, Language B, and the unidentified languages, and running the glyph counts using the "Find" function in Microsoft Word. This analysis is summarized in Table 5.3.

We observe here that some glyphs are noticeably more frequent in one language than another. As a rough and ready interpretation of this table, we have highlighted those instances when there are differences of at least 2 percentage points. These include the following:

- The glyph o {o} is 4.6 percentage points more frequent in A than in B.
- The glyph c {c} is 3.0 percentage points more frequent in B than in A.

We guess that differences such as these caught Currier's expert eye. We imagine that the frequencies of the glyph o {o} in particular would have informed his call as to whether a page was in Language A or B.

As illustrated in Table 5.3, the divergences in language-specific frequencies, vis-à-vis the manuscript as a whole, were within the range +2.9 percentage points to –2.8 percentage points. We can reasonably pose the question: do these differences point to a different language, or are they within the normal variations of documents in a single language?

We recall that the Wilfrid Voynich purchased the manuscript from a Jesuit college near Frascati in Italy. If we believe that the manuscript was derived from a natural language or languages, then by the distance-decay theory,[6] we should look for those languages in concentric circles centered on Italy. The farther from Italy, the less probable are the languages. The prime candidate is the medieval Italian language itself, or else Renaissance Latin. The next circle of languages, remembering the manuscript's putative travels through Vienna and Prague, and its unmistakably French annotations, could be French or other Romance languages, as well as German and Bohemian. The circle beyond that could include Greek and the Balkan[7] and Slavic[8] languages.

Table 5.3. Voynich manuscript, v101 transliteration: counts of the eighteen most frequent glyphs, plus spaces and uncertain spaces, by Currier language

glyph	v101	Currier language: all		A		B	
glyphs + spaces		194,286	100.0%	55,548	100.0%	123,591	100.0%
space		32,879	16.9%	9,428	17.0%	20,842	16.9%
"uncertain space"		2,460	1.3%	621	1.1%	1,621	1.3%
glyph total		158,947	81.8%	45,499	81.9%	101,128	81.8%
	o	25,043	12.9%	8,643	15.6%	13,580	11.0%
o	part of †o	5,265	2.7%	1,080	1.9%	4,095	3.3%
	other	19,778	10.2%	7,563	13.6%	9,485	7.7%
9	9	17,436	9.0%	4,439	8.0%	11,320	9.2%
a	a	14,496	7.5%	3,538	6.4%	9,248	7.5%
c	c	13,750	7.1%	2,638	4.7%	9,541	7.7%

ᴑᴄ	1	11,052	5.7%	3,921	7.1%	6,233	5.0%
ᶍ	e	10,652	5.5%	3,015	5.4%	6,610	5.3%
ᵹ	8	10,343	5.3%	2,841	5.1%	6,767	5.5%
ᶋ	h	9,919	5.1%	2,312	4.2%	6,718	5.4%
ᶎ	y	6,717	3.5%	2,139	3.9%	3,853	3.1%
ᶋ	k	6,118	3.1%	1,651	3.0%	3,610	2.9%
	4	5,408	2.8%	1,130	2.0%	4,219	3.4%
ᴏ	part of ᵗᴏ	5,265	2.7%	1,080	1.9%	4,095	3.3%
	other	143	0.1%	50	0.1%	124	0.1%
ᵢᵢᵑ	m	4,112	2.1%	1,419	2.6%	2,409	1.9%
ᵭ	2	3,359	1.7%	1,072	1.9%	2,116	1.7%
ᴄᴄ	C	2,844	1.5%	450	0.8%	2,099	1.7%
ᵹ	7	2,714	1.4%	332	0.6%	2,100	1.7%
ᶐ	s	2,702	1.4%	1,042	1.9%	1,209	1.0%
ᵢᵑ	n	1,770	0.9%	273	0.5%	1,448	1.2%
ᵱ	p	989	0.5%	349	0.6%	557	0.5%

Source: author's analysis based on Currier's allocation of folios to languages.

over 2 percentage points above average for manuscript

1-2 percentage points above average for manuscript

1-2 percentage points below average for manuscript

over 2 percentage points below average for manuscript

We therefore looked at the frequencies of common letters in medieval Italian, and in several other medieval European languages, as illustrated in Table 5.4.

From this table we can see that in these languages, the frequencies of the single most common letter could diverge by up 11 percentage points; however, the next most common letters were not very divergent in frequency. In this respect, the languages of the Voynich manuscript are

unusual, with four of the five most common glyphs diverging in frequency by more than 2 percentage points.

If on the other hand we look at documents in a single language, we might expect the letter frequencies to be less divergent. We put this idea to the test by comparing a large corpus of medieval Italian with a specific document in medieval Italian; namely, Dante Alighieri's *La Divina Commedia*.

Dante started writing *La Divina Commedia* around 1308 and completed it in 1320. It preceded the production of the Voynich parchment by up to a century. It is an epic poem of 14,233 lines and 403,058 characters,[9] written in the Tuscan Italian of Dante's time. Dante's Italian is not far removed from modern Italian; we guess that a twenty-first-century Italian speaker could understand it at least as easily as a modern English speaker could understand Shakespeare.

We downloaded the text of *La Divina Commedia* from the Project Gutenberg website, imported it into Microsoft Word, counted the common letters, and compared their frequencies with those in the OVI corpus of medieval Italian. The results are shown in Table 5.5. Here we can see that in *La Divina Commedia*, the frequencies of the common letters do not diverge by more than +0.7 or –1.0 percentage points from those in medieval Italian as a whole.

Table 5.4. Frequencies of the ten most common letters in selected medieval European languages

Old Albanian		Bohemian		Middle English		Middle French		Early New High German		Italian, medieval		Latin, medieval		ave
E	22.0%	I	12.7%	E	16.5%	E	17.4%	E	18.8%	E	11.6%	E	11.1%	15
T	9.8%	E	11.9%	I	7.9%	S	8.3%	N	10.2%	A	10.5%	I	10.0%	9.
I	9.1%	A	6.8%	T	7.6%	A	7.8%	I	8.0%	I	9.8%	U	8.8%	8.
N	7.1%	S	6.8%	N	7.5%	N	7.1%	R	7.2%	O	9.4%	T	8.0%	7.
H	6.4%	O	5.8%	O	7.2%	I	7.0%	S	6.3%	N	6.6%	S	7.7%	6.
A	6.1%	Z	5.6%	S	6.4%	R	6.5%	T	6.2%	R	6.5%	A	7.1%	6.
U	4.5%	T	4.9%	H	6.1%	T	6.2%	A	5.2%	T	5.9%	N	6.3%	5.
O	4.1%	N	4.3%	A	5.1%	O	6.1%	D	4.9%	L	5.8%	M	5.9%	5.

ſ	4.1%	W	4.1%	D	4.7%	U	5.2%	L	4.6%	S	5.6%	O	5.9%	4.5
M	3.9%	R	3.7%	R	4.5%	L	4.9%	H	4.2%	C	5.2%	R	4.8%	4.4

Source: author's analysis based on Gjon Buzuku, *Missale* (Albanian); *Dalimilova kronika* (Bohemian); Marsha Lynn Moreno (English); *La Farce de Maistre Pathelin* (French); *Der Ackerman aus B ö hmen* (German); OVI corpus (Italian); LatinISE medieval subcorpus (Latin). Note: frequencies are not case-sensitive, and distinguish letters with accents.

frequency is over 2 percentage points above average

frequency is over 2 percentage points below average

Table 5.5. Frequencies of the most common letters in medieval Italian, and in La Divina Commedia by Dante Alighieri

	medieval Italian (OVI)	*La Divina Commedia*	
letter	percent	count	percent
all	100.0%	403,058	100.0%
E	12.6%	46,680	11.6%
A	10.0%	42,580	10.6%
I	9.4%	39,627	9.8%
O	9.3%	37,778	9.4%
N	6.8%	26,553	6.6%
T	5.4%	22,861	5.7%
R	6.5%	26,024	6.5%
L	6.2%	23,499	5.8%
S	5.0%	22,665	5.6%
C	4.8%	20,887	5.2%
D	4.8%	15,040	3.7%
U	3.7%	13,485	3.3%
P	2.9%	11,331	2.8%
M	2.9%	12,006	3.0%
V	1.8%	8,169	2.0%

Source: analysis by author based on OVI corpus (medieval Italian), and text of *La Divina Commedia* (with modern spelling) from Project Gutenberg.

Notes: frequencies are based on letter counts without case sensitivity.

So, at least from this example, the divergences that we have seen in Currier's Languages A and B do seem more like flags of a difference in language than indications of variation between documents.

To give an example: o {o} and 9 {9} are the first and second most common glyphs in the v101 transliterations of both Languages A and B. But cc {1}, which is the third most common in A, drops to fifth place in B. This is as if in English the letter "a" (which is generally agreed to be the third most frequent letter) became the fifth most frequent. This is a greater variation than that between American English and British English; it is more like the difference between modern English and medieval English.

ZIPF'S LAW

This may be as good a place as any to continue the analysis that we started in Chapter 4: to see to what extent the individual Voynich languages conform to Zipf's Law. The results are summarized in Table 5.6 and figure 5.2.

From Table 5.6 we can see that Languages A and B do indeed look like substantially different methods of communication. Of the top twenty words in Language B, eleven do not appear at all in the top twenty of Language A. If Languages A and B were both variants of American English, it is as if Language B used the word "a" and Language A did not.

Table 5.6. Voynich manuscript, v101 transliteration: the twenty most common "words" in Languages A and B

	Voynich manuscript				Language A				La	
rank	glyphs	v101	count	% of 1st "word"	glyphs	v101	count	% of 1st "word"	glyphs	v1(
1	ꞩa ıı∂	8am	739	100.0%	ꞩa ıı∂	8am	429	100.0%	o9	o(

1	ᾰ ιι ᾰ	am	739	100.0%	ᴄᴏᵩ	1oe	276	64.3%	ᾰ ιι ᾰ	ar
3	ᴏᵩ	oe	710	96.1%	ᾰ	s	244	56.9%	ᾰ ᾰ	a
4	ᾰ ᾰ	ay	489	66.2%	ᴄᴏ ᾰ	1oy	185	43.1%	ᴄᴄ8 9	1c8
5	ᴏ ᾰ	oy	440	59.5%	ᴏᵩ	oe	161	37.5%	ᴏ ᾰ	o
6	ᾰ	s	432	58.5%	ᾰ ιι ᾰ	am	140	32.6%	8ᾰ ιι ᾰ	8a
7	ᴄᴏᵩ	1oe	380	51.4%	9	9	137	31.9%	ᴄᴄ9	1c
8	ᴄᴄ9	1c9	375	50.7%	ᴏ ᾰ	oy	131	30.5%	ᾰᵩ	ae
9	ᾰᵩ	ae	361	48.8%	ᴄ9	19	127	29.6%	†ᴏ ᾰ ιι ᾰ	4oh
10	ᴄᴄ8 9	1c89	360	48.7%	8 9	89	119	27.7%	9	9
11	9	9	315	42.6%	ᴋ9	K9	92	21.4%	ᵩ	e
12	8 9	89	294	39.8%	ᴄ̂ᴏᵩ	2oe	91	21.2%	†ᴏ ᾰ ιι ᾰ	4oh
13	8ᾰ ᾰ	8ay	290	39.2%	8ᾰ ᾰ	8ay	89	20.7%	ᴄ̂ᴄ8 9	2c8
14	ᾰ	y	278	37.6%	ᴄᴄ9	1c9	88	20.5%	†ᴏ ᴄᴄ9	4oh
15	ᵩ	e	270	36.5%	ᴄ̂ᴏ	2o	86	20.0%	ᾰ	y
16	†ᴏ ᾰ ιι ᾰ	4ohan	264	35.7%	ᴏ	o	86	20.0%	8ᾰ ᾰ	8a
17	†ᴏ ᾰ ιι ᾰ ιι	4oham	233	31.5%	8ᾰ ιι	8an	82	19.1%	ᴄᴄ8 9	1c7
18	†ᴏ ᴄᴄ9	4ohC9	232	31.4%	ᾰ	y	80	18.6%	ᾰ	s
19	8ᾰᵩ	8ae	222	30.0%	8ᾰᵩ	8ae	80	18.6%	†ᴏ ᴄ8 9	4oh
20	ᴄᴏ ᾰ	1oy	217	29.4%	ᴄᴏ	1o	79	18.4%	ᾰ ιι	a

Source: author's analysis based on Currier's allocation of folios to languages. "Words" are defined as glyph strings delimited by spaces, uncertain spaces or line breaks. "Word" counts are by author from WordCounter, and are case-sensitive by reference to v101.

only in top 20 of Language A

only in top 20 of Language B

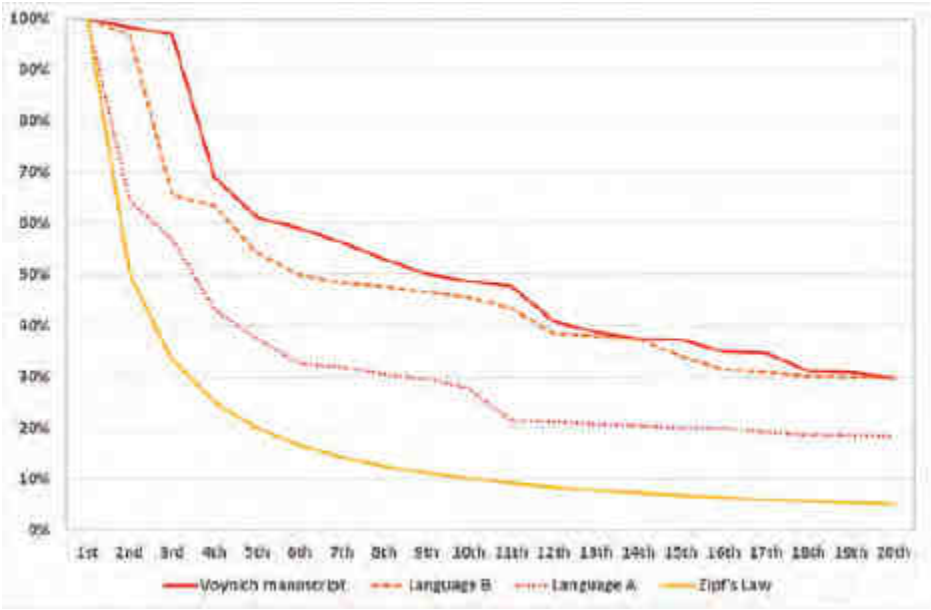


Figure 5.2. Voynich manuscript and Currier language sections: frequencies of the twenty most common words. *Source: author's analysis*

On the other hand, we cannot conclude that Languages A and B are different languages in the modern sense, since they share many words, including the top three in Language B: ol {oe}, a ıı̃ {am} and aʔ {ay}. If Language B were modern American English, Languages A and B would have to be languages which use the words "the," "of" and "and." Certainly, we can find "the" in Danish and Swedish, even in French (as "thé," meaning "tea"); "of" exists in Dutch, meaning "or"; there is a word "and" in Danish, Norwegian, and Swedish, meaning "duck." But we would be hard-pressed to imagine a document in any modern language in which "and" and "duck" are both relatively common.

In short, Language A seems to have a significantly different vocabulary from Language B, but they are not different languages. If we may use a modern analogy: the two languages differ from each other in the way that American English differs from British English, but more so. While American English has "sidewalk" for "pavement" and "trunk" for "boot," we need to imagine a variant of American or British English which differs even in the common words.

Returning to Zipf's Law: from figure 5.2 we can see that the Language B section is the more divergent from Zipf's Law. Language A is in closer

conformity to Zipf's Law; but in this author's view, that does not make it a different language from B.

THE HANDS

This brings us to the "hands" whose owners left the imprint of their inalienable personalities on the Voynich manuscript. Currier had this to say:

> … in the middle of [the "pharmaceutical" section] … there are six pages (two folios, folded so that there are three pages on each) which show a very obvious difference in hand: cramped, slanted, having quite a different character, very obvious even to the untrained eye.
>
> … The recipe section … contains only one folio on which the writing differs noticeably to the eye from that on other folios …
>
> … There were, in fact, only two "languages," but perhaps as many as eight or a dozen different identifiable hands.[10]

Within a few lines of his presentation, he modified this estimate:

> … we have a total of something like five or six to seven or eight different identifiable hands in the manuscript. This gives us a total of two languages and six to eight scribes (copyists, encipherers, call than what you will).[11]

In his appendix table he explicitly listed seven "hands" as follows: 1, 2, 3, 4, 5, X, and Y. Regarding the six pages of the "pharmaceutical" section where the handwriting had seemed to him so egregiously different, in the end he left the "hand" unidentified.

By Currier's reckoning, therefore, there were only two major scribes:

- the owner of “hand 1” who wrote most of the “herbal” section, plus the first page of the manuscript, which is text-only and which we are tempted to see as a key to the “herbal” section
- the owner of “hand 2” who wrote a big chunk of the “herbal” section and the whole of the “biology / balneology” section

Table 5.7. The scribes who wrote each thematic section of the Voynich manuscript

thematic section	number of pages, by scribe					
	all	scribe 1	scribe 2	scribe 3	scribe 4	scribe 5
herbal	129	95	20	8	0	6
astronomy	8	0	0	0	8	0
cosmology	11	1	0	0	9	0
astrology	12	0	0	0	12	0
biology / balneology	19	0	19	0	0	0
pharmaceutical	16	16	0	0	0	0
text with marginal stars	25	0	1	25	0	0
text only	7	1	1	1	3	1
total	227	113	41	34	32	7

Source: analysis by author based on Davis, Lisa Fagin. “How Many Glyphs and How Many Scribes? Digital Paleography and the Voynich Manuscript”. *Manuscript Studies* : Vol. 5, Issue 1 (Spring 2020), Article 6.

while the owners of the other “hands” contributed a few pages here and there.

In the Voynich 2022 conference, Dr. Lisa Fagin Davis expanded on Currier’s identification of the “hands” (which she called “scribes”). She filled in all of Currier’s omissions: that is, she identified a scribe for all the pages for which Currier had not done so. She proposed that there had been five scribes (versus Currier’s eight or more). The sections of text which she ascribed to Scribes 1 and 2 were, with very few exceptions, the same as those defined by Currier. She agreed with Currier that Scribe 1 had used Language A and Scribe 2 had used Language B.

She made the additional proposal that the scribes whom she numbered 3, 4, and 5 had used Language B. Versus Currier’s estimates, this proposal

allocated 28 additional pages of the Voynich text to Language B. If we accept this proposal, it creates a slightly larger database whereby we can test the statistical properties of Language B against those of natural languages.

Dr. Davis found two instances where two scribes had worked on the same folio. In folio f057, she saw the hand of Scribe 5 on the recto and that of Scribe 1 on the verso. In folio f112 recto, she identified the first twelve lines as the hand of Scribe 2 and the rest as that of Scribe 3 (both writing in Language B).

We cannot easily reconstruct Currier's or Davis's analysis of the "hands." There is no computer-readable version of the manuscript which preserves the original glyphs; therefore, we cannot say what shapes or styles signalled for them a specific scribe, and which they thought were merely variations in the handwriting of an individual. We know which pages they assigned to which "hands," and with modern OCR software we could scan every one of those pages and identify the characters that differed from one scribe to another. We could even create mathematical expressions of those differences. But that is a task beyond the scope of this book; and even if this task were done, we could not expect it to elucidate the meaning of the text, if meaning is there to be found.

THE THEMATIC SECTIONS

Capt. Currier's identification of the languages implied that in each of the thematic sections that he studied, Language A or B predominated, or was the sole language. If his Languages A and B were truly different, we should then expect the glyph frequencies to differ from one thematic section to another.

This was an easy hypothesis to test: it was a matter of splitting the text into seven chunks corresponding to the seven sections, importing each chunk into Microsoft Word, applying the "find" function, and counting the glyphs in each chunk (at least the most common ones). The results are shown in Table 5.8.

Table 5.8. The Voynich manuscript, v101 transliteration: counts of the ten most common glyphs, by thematic section

glyph	v101	thematic section: all		astronomy		cosmology		astrology / zodiac	
characters		194,286	100.0%	4,448	100.0%	10,731	100.0%	7,073	100.0%
spaces		35,339	18.2%	626	14.1%	1,898	17.7%	1,032	14.6%
glyph total		158,947	81.8%	3,822	85.9%	8,833	82.3%	6,041	85.4%
o	o	19,778	10.2%	644	14.5%	1,329	12.4%	1,106	15.6%
9	9	17,436	9.0%	454	10.2%	925	8.6%	610	8.6%
a	a	14,496	7.5%	309	6.9%	994	9.3%	710	10.0%
c	c	13,750	7.1%	392	8.8%	765	7.1%	675	9.5%
ch	1	11,052	5.7%	295	6.6%	521	4.9%	284	4.0%
s	e	10,652	5.5%	174	3.9%	503	4.7%	475	6.7%
d	8	10,343	5.3%	253	5.7%	564	5.3%	192	2.7%
f	h	9,919	5.1%	222	5.0%	458	4.3%	309	4.4%
r	y	6,717	3.5%	167	3.8%	476	4.4%	248	3.5%
f	k	6,118	3.1%	165	3.7%	447	4.2%	404	5.7%

		balneology / biology		herbal		pharmaceutical		text	
characters		32,842	100.0%	53,682	100.0%	13,061	100.0%	72,449	100.0%
spaces		5,876	17.9%	10,052	18.7%	2,297	17.6%	13,558	18.7%
glyph total		26,635	81.1%	43,084	80.3%	10,608	81.2%	57,708	79.7%
o	o	2,217	6.8%	6,514	12.1%	2,004	15.3%	5,964	8.2%
9	9	3,464	10.5%	4,957	9.2%	989	7.6%	6,037	8.3%
a	a	1,800	5.5%	3,549	6.6%	760	5.8%	6,373	8.8%
c	c	2,798	8.5%	2,738	5.1%	1,028	7.9%	5,354	7.4%
ch	1	1,337	4.1%	4,061	7.6%	630	4.8%	3,924	5.4%
s	e	2,272	6.9%	2,401	4.5%	863	6.6%	3,964	5.5%
d	8	1,996	6.1%	3,088	5.8%	574	4.4%	3,676	5.1%
f	h	1,887	5.7%	2,453	4.6%	651	5.0%	3,939	5.4%
r	y	831	2.5%	2,040	3.8%	507	3.9%	2,448	3.4%

ſſ	k	831	2.5%	1,799	3.4%	218	1.7%	2,254	3.1%

Source: analysis by author; thematic sections as per Capt. Prescott Currier, *Voynich manuscript, some notes and observations*, in D'Imperio, Mary (1976), Appendix A. Counts of spaces include "uncertain spaces". Counts of ○

over 2 percentage points above average for manuscript

over 2 percentage points below average for manuscript

Our reading of this table is that, certainly, there are some differences in the glyph frequencies from one section to another. For example, ○ {o} is more common in the astronomy, zodiac, and pharmaceutical sections; ɑ {a} is most frequent in the zodiac section and most scarce in the biology section. But the differences are generally not great, nor are they pervasive. For each of the seven sections as defined here, the differences between the glyph frequencies in the section and those in the manuscript as a whole averages less than 0.5 percentage points.

We are tempted, therefore, to wonder whether the statistical differences between Language A and Language B really are flags of two different languages, or whether they are simply indicators of two different source documents in a single language.

Decades after Capt. Currier's analysis, researchers Marcelo Montemurro and Damián Zanette would return to the thematic sections, but with no prior assumption as to whether they were in different languages or not. We shall report on their work in Chapter 7.

THE FUNCTION OF THE LINE

In any case, Currier's best was yet to come. He went on to tell his audience that:

> … the line is a functional entity in the manuscript on all those pages where the text is presented linearly…. *the line itself is a functional unit*.[12] [our emphasis]

For our understanding of the Voynich manuscript, this statement calls for deep consideration. Currier was saying that a line of the manuscript had some property that allowed it to stand alone. If he did not quite say so, he was implying that if there was meaning in the manuscript, then a single line could possess meaning, independently of other lines.

Let us consider, for example, the second line of the American Declaration of Independence:

> … assume among the powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they…[13]

When we read this line, we know immediately that it does not stand alone. We know that it is part of a larger text and that it is preceded and followed by other text, in combination with which, and not otherwise, it has meaning.

Let us consider the second line of Marc Antony's address to the Romans in Shakespeare's *Julius Caesar*:

> I come to bury Caesar, not to praise him.[14]

This line surely has a context; but it stands alone. Currier was saying that the lines of the Voynich manuscript, in some respect, possessed this property. In some way, they were self-sufficient.

Currier gave three reasons why the lines (he meant, specifically, horizontal lines as distinct from circular text, radial text or stand-alone labels) had some function in their own right:

> The frequency counts of the beginnings and endings of lines are markedly different· from the counts of the same characters

> internally. There are, for instance, some characters that may not occur initially in a line. There are others whose occurrence as· the initial syllable of the first "word" of a Iine is about one hundredth of the expected.
>
> … The ends of lines contain what seem to be, in many cases, meaningless symbols: little groups of letters which don't occur anywhere else, and just look as if they were added to fill out the line to the margin.
>
> … I have three computer runs of the herbal material and of the biological material. In all of that, which is almost 25,000 "words," there is <u>not one single case</u> [his emphasis] of a repeat going over the end of a line to the beginning of the next; not one.[15]

In the last point, it is not clear what Currier meant by a "repeat," but his implication was clear: a given line is not a continuation of the previous line. In modern parlance, the text does not "wrap."

Extraordinarily, Currier let his discoveries rest there. He would only say, in conclusion:

> I have no interpretations of them, by the way. I have no solutions. All I know is that they are significant—and damn significant…. They are there, and they are very definite.[16]

We can pick up this ball and run with it. Equipped as we are with twenty-first-century computing power, we were able to reconstruct in minutes what might have taken Currier weeks or months to prepare. The process was as follows.

Again, we had the full text of the Voynich manuscript in the v101 alphabet; this time Microsoft Excel was the preferred format. In Excel it was straightforward to extract the first and last characters of cells, using the LEFT and RIGHT functions. It was then a matter of counting the occurrences of each of the Voynich glyphs (for convenience, sticking to the twenty-seven most common ones). Here the COUNTIF function would have been the first choice, were it not that this function is not case sensitive,

and the v101 alphabet includes, among its representations of the common glyphs, both c and C; and both h and H. We are grateful to Dave Bruns and Lisa Bruns for supplying the case-sensitive counting formula.[17] The results are summarized in Table 5.9.

Table 5.9. The Voynich manuscript, v101 transliteration: frequencies of the twenty-seven most common glyphs, in the whole manuscript, and as the initial and final glyphs of horizontal lines

glyph	v101	whole manuscript		as initial glyph of horizontal line		as final glyph of horizontal line	
		count	percent	count	percent	count	percent
all	all	158,947	100.0%	4,129	100.0%	4,129	100.0%
o	o	25,043	15.8%	547	13.2%	52	1.3%
9	9	17,436	11.0%	600	14.5%	1,585	38.4%
a	a	14,496	9.1%	11	0.3%	7	0.2%
c	c	13,750	8.7%	5	0.1%	2	0.0%
1	1	11,052	7.0%	129	3.1%	1	0.0%
e	e	10,652	6.7%	69	1.7%	456	11.0%
8	8	10,343	6.5%	595	14.4%	99	2.4%
h	h	9,919	6.2%	92	2.2%	0	0.0%
y	y	6,717	4.2%	9	0.2%	317	7.7%
k	k	6,118	3.8%	400	9.7%	3	0.1%
4	4	5,408	3.4%	522	12.6%	0	0.0%
m	m	4,112	2.6%	0	0.0%	400	9.7%
2	2	3,359	2.1%	163	3.9%	0	0.0%
C	C	2,844	1.8%	0	0.0%	0	0.0%
7	7	2,714	1.7%	37	0.9%	7	0.2%
s	s	2,702	1.7%	418	10.1%	133	3.2%
n	n	1,770	1.1%	0	0.0%	167	4.0%
p	p	989	0.6%	0	0.0%	591	14.3%
K	K	910	0.6%	17	0.4%	2	0.0%
g	g	879	0.6%	340	8.2%	0	0.0%

ᓂᐦᐨ	H	866	0.5%	0	0.0%	0	0.0%
ᓇ	A	768	0.5%	11	0.3%	7	0.2%
ᐩ	j	564	0.4%	29	0.7%	1	0.0%
ᓵ	3	516	0.3%	15	0.4%	0	0.0%
ᔑ	z	511	0.3%	0	0.0%	11	0.3%
ᓴ	5	409	0.3%	7	0.2%	1	0.0%
ᓀ	(	402	0.3%	7	0.2%	22	0.5%
rare glyphs	@xxx;	341	0.2%	24	0.6%	0	0.0%

Source: analysis by author. Glyph counts exclude spaces and uncertain spaces.

position-specific frequency exceeds manuscript frequency by over 2 percentage points

position-specific frequency is below manuscript frequency by over 2 percentage points

From this table, we can see that when Currier called his findings “damn significant,” he was being modest. The anomalies fairly leap off the page. Nearly every one of the eighteen most common glyphs has a substantially different frequency (and in several cases, is entirely absent) if it is the first or last glyph of a horizontal line. Eight glyphs, namely ᐧ {9}, ᐃ {8}, ᐠ {o}, ᐟ {4}, ᐢ {s}, ᐩᐨ {k}, ᐩ {g} and ᓴ {2}, account for 87 percent of the initial glyphs of the horizontal lines. Seven glyphs, but generally not the same ones, namely ᐧ {9}, ᐨ {p}, ᐪ {e}, ᑊᐁ {m}, ᐣ {y}, ᑊᐁ {n} and ᐢ {s}, account for 88 percent of the final glyphs. The glyph ᐧ {9} stands out both as a very common initial glyph of the line, and as a very common final glyph of the line.

At this point we might be tempted to say that these thirteen glyphs (and maybe all of the glyphs), if they have any meaning, change or even lose that meaning when they are at the beginning or end of a line. Possibly the last few glyphs of every line are meaningless fillers, as Currier suspected. Possibly the same is true of the first few glyphs. We are starting to imagine the authors or producers of the manuscript giving instructions to the scribes, along the following lines (no pun intended):

> Here's the line for you to transcribe into the secret alphabet. But before you start the transcription, put a few random glyphs at the beginning. And when you reach the end, add a few more until you reach an illustration, or the right margin.

If we were extraterrestrials with no prior knowledge of the English language, we might be similarly puzzled to read the following rendition of Percy Shelley's "Ozymandias":

> rslefI met a traveler from an antique landymrqy
> njifkWho said: two vast and trunkless legs of stonemufmu
> aopkfStand in the desert. Near them, on the sandwcjzw[18]

But before we tread this perilous path with regard to the Voynich manuscriupt, let us investigate what letter frequencies we might expect in a comparable text, in a known language, in which the lines are somewhat independent of each other. Even better, let us look at a text that is not too distant in time from the Voynich parchment (which, as we recall, was made from calf hides dated between 1308 and 1458). For this purpose, let us look again at *La Divina Commedia*.

In our digital copy of *La Divina Commedia*, we removed all punctuation that occurred at the beginnings and ends of lines and extracted the initial and final letters of each line. We then used the COUNTIF function to count the letter frequencies. The results are shown in Table 5.10, including a comparison with medieval Italian.

We see in *La Divina Commedia* a divergence from the Italian language corpus that is not unlike what we have seen in the Voynich manuscript. All of the most common letters have unusual frequencies when they occur at the beginning or end of a line. The initial letter of the line is, far more often than in other occurrences, "c," "d," "m," "s," "p," or "q." The terminal letter of the line is, in 98 percent of the cases, "a," "o," "e," or "i" as we would expect from the Italian word endings in the masculine and feminine gender, singular and plural. Most of the common letters also have unusual frequencies when they occur as the second letter, or the second from last letter, in the line. So the lines of the Voynich manuscript are no more bizarre

in the frequencies of their initial and final glyphs than are the lines of *La Divina Commedia*.

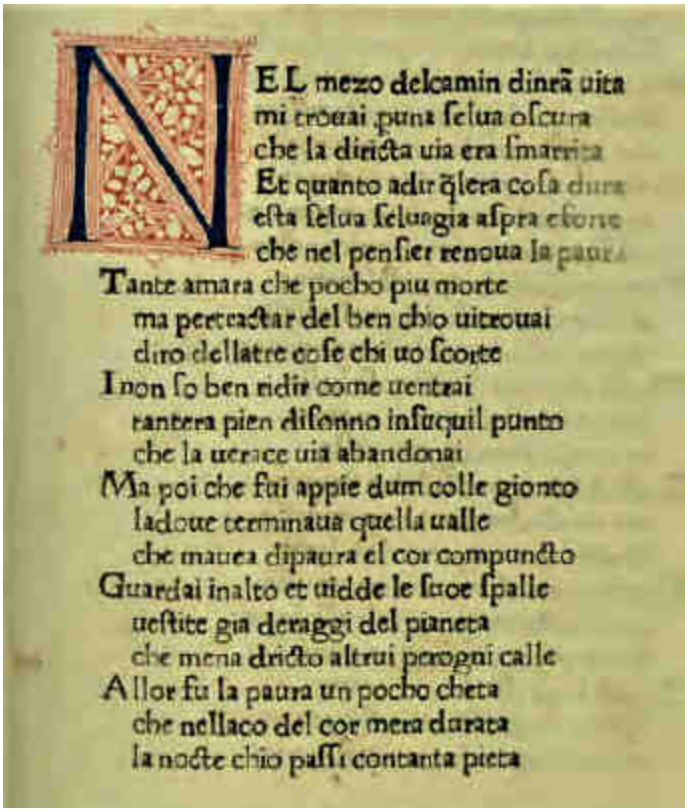

NEL mezo delcamin dinrã uita
mi troual puna ſelua oſcura
che la dirícta uia era ſmarrita
Et quanto adir q̃lera coſa dura
eſta ſelua ſeluagia aſpra eforte
che nel penſier renoua la paura
Tante amara che pocho piu morte
ma pertractar del ben chio uitrouai
dirò dellaltre coſe chi uo ſcorte
I non ſo ben ridir come uentrai
tantera pien diſonno inſuquil punto
che la uerace uia abandonai
Ma poi che fui appie dum colle gionto
ladoue terminaua quella ualle
che mauea dipaura el cor compuncto
Guardai inalto et uidde le ſuoe ſpalle
ueſtite gia deraggi del pianeta
che mena dricto altrui perogni calle
Allor fu la paura un pocho cheta
che nellaco del cor mera durata
la nocte chio paſſi contanta pieta

Figure 5.3. *La Divina Commedia* by Dante Alighieri: the first twenty-one lines. Every one of these lines ends in “a,” “e,” “i,” or “o.” *Public domain*

If we are considering medieval Italian as a candidate language, we could now begin to consider the hypothesis that the Voynich final glyphs correspond to Italian final letters.

For example, matching up the most-common line-ending glyphs in the Voynich manuscript and the most-common line-ending letters in *La Divina Commedia*, we could hypothesize the following:

- that ⟨o⟩ {9} could indicate the feminine singular “a”
- that ⟨ꝑ⟩ {p}, or possibly just the last part ⟨ꞵ⟩ {*}, could indicate the masculine singular “o”
- that ⟨ꬲ⟩ {e} could indicate the feminine plural “e”

- that 𐤌𐤌𐤍 {m}, or possibly the last part, 𐤍 {N}, could indicate the masculine plural "i,"

or some permutation of these mappings.

Table 5.10. Letter frequencies in medieval Italian, and in the first and last characters of the lines in La Divina Commedia

letter	medieval Italian precent	*La Divina Commedia* as 1st letter of line		as 2nd letter of line		as 2nd last letter of line		as last letter of line	
		count	percent	count	percent	count	percent	count	percent
all	100.0%	14367	100.0%	14367	100.0%	14367	100.0%	14367	100.0%
E	12.5%	1678	11.7%	1981	13.8%	193	1.3%	3481	24.2%
A	9.9%	573	4.0%	1584	11.0%	140	1.0%	4413	30.7%
I	9.3%	457	3.2%	1610	11.2%	1173	8.2%	2280	15.9%
O	9.3%	481	3.3%	2308	16.1%	96	0.7%	3971	27.6%
N	6.7%	691	4.8%	470	3.3%	1572	10.9%	6	0.0%
R	6.5%	201	1.4%	455	3.2%	1490	10.4%	3	0.0%
L	6.1%	995	6.9%	194	1.4%	950	6.6%	0	0.0%
T	5.4%	516	3.6%	38	0.3%	3271	22.8%	0	0.0%
S	4.9%	1079	7.5%	71	0.5%	1428	9.9%	3	0.0%
D	4.7%	1304	9.1%	137	1.0%	1085	7.6%	0	0.0%
C	4.7%	2381	16.6%	29	0.2%	795	5.5%	0	0.0%
U	3.7%	99	0.7%	1222	8.5%	212	1.5%	1	0.0%
M	2.9%	710	4.9%	45	0.3%	522	3.6%	0	0.0%
P	2.9%	1322	9.2%	44	0.3%	128	0.9%	0	0.0%
V	1.8%	432	3.0%	57	0.4%	453	3.2%	1	0.0%
G	1.8%	221	1.5%	25	0.2%	264	1.8%	0	0.0%
H	1.5%	10	0.1%	1363	9.5%	124	0.9%	2	0.0%
F	1.3%	358	2.5%	3	0.0%	27	0.2%	0	0.0%
Q	0.9%	710	4.9%	1	0.0%	0	0.0%	0	0.0%
B	0.8%	113	0.8%	8	0.1%	110	0.8%	0	0.0%

Z	0.4%	1	0.0%	0	0.0%	216	1.5%	0	0.0%
È	0.4%	32	0.2%	3	0.0%	7	0.0%	3	0.0%
Ò	0.3%	0	0.0%	4	0.0%	6	0.0%	3	0.0%
Ì	0.3%	0	0.0%	289	2.0%	19	0.1%	9	0.1%
À	0.3%	0	0.0%	64	0.4%	0	0.0%	0	0.0%
É	0.1%	0	0.0%	95	0.7%	2	0.0%	4	0.0%
Ù	0.1%	0	0.0%	9	0.1%	14	0.1%	2	0.0%

Source: Author's analysis based on OVI corpus (medieval Italian), and text of *La Divina Commedia* from Project Gutenberg

Notes: frequencies are based on letter counts without case sensitivity. Initial and final punctuation was removed from *La Divina Commedia*.

over 2 percentage points higher than in medieval Italian corpus

over 2 percentage points lower than in medieval Italian corpus

However, the mapping is not brilliant; for example, the frequencies of the line-final glyphs have only a 75.3 percent correlation with those of the line-ending letters in *La Divina Commedia*. This is largely because the line endings of *Commedia* are dominated by four letters, whereas the line endings of the Voynich manuscript are more diverse. We need to look at other medieval texts with a line-by-line structure.

Gjon Buzuku's *Meshari* or *Missale* (of which more in Chapter 13), published in 1555, is the earliest extant text in Old Albanian. It is a work of prose, but not narrative prose; it is a translation of the main parts of the Catholic liturgy. As such, it is highly structured and repetitive (like the Voynich manuscript). It has approximately 16,800 lines. On 4,199 lines, a word wraps from the line to the next, with a hyphen to denote the break in the word. If, purely hypothetically, we imagine *Meshari* as a source document for the Voynich manuscript, we could suppose that the Voynich scribes mapped the hyphen to a glyph; if so, inevitably that glyph would be the ubiquitous final ୨ {9). Indeed, if we do the mathematics of such a mapping, the correlation works out at 89.2 percent,[19] which is an improvement on *La Divina Commedia*.

Better still: in addition to the hyphen, *Meshari* has several punctuation marks—notably the period and the colon—which often appear at the end of a line. If we elaborate our Voynich scenario with the idea that all final punctuation marks mapped to a single glyph, again that glyph would be the final ⟨9⟩ {9}; and the correlation would improve to 95.7 percent.[20]

In such a scenario, we can see a possible solution to the mystery of the ⟨9⟩ {9). It may be so pervasive as the final glyph and as the initial glyph of a "word," and as the final glyph of a line, because it does not represent a letter, but a punctuation mark. What applies to the ⟨9⟩ {9} may also apply to the ubiquitous ⟨o⟩ {o}, to the penultimate ⟨8⟩ {8} and ⟨c⟩ {c}, and to the extraordinary ⟨to⟩ {4o}.

In a later chapter we shall present some tests of this idea. It could be a productive field of research for any interested reader, to examine the frequencies of both letters and punctuation marks in medieval documents and assess to what extent these frequencies line up with those of the Voynich glyph set.

THE VOYNICH MANUSCRIPT AS POETRY

Our investigation of *La Divina Commedia* and *Meshari* prompts us to suggest what Currier could have said but chose not to say: that since the lines of horizontal text in the Voynich manuscript are independent of each other in some way, it follows that the manuscript is not a work of prose.

One possible interpretation is that the manuscript is a book of lists, or a catalogue. But if so, its entries are unusual, in that, on any given page, they tend to be of equal length. Generally, they start at the left margin and end at or near the right margin, subject to the meticulous avoidance of any intervening illustration. That would take us back to Currier's "junk filler" hypothesis.

If the manuscript is a liturgy along the lines of *Meshari*, then we need to search for the names of the Deity and of the saints and prophets.

Another interpretation is that the manuscript is poetry. To this idea, we shall return.

POSTSCRIPT: PUNCTUATION

Capt. Currier clearly was troubled by the proliferation of glyphs which, in his perception, had no meaning. However, in his presentation to the 1976 seminar, he never mentioned punctuation. In modern English and in most Indo-European languages, punctuation marks are the predominant characters in any text which convey information to the reader but have no intrinsic meaning.

For example, a period tells us that a sentence has ended. A comma tells us that a phrase has ended, and a new one is beginning. An exclamation mark tells us that the narrator (or a character within the narration) is shocked or surprised. And so on.

We took the opportunity to pursue Capt. Currier's idea of meaningless glyphs, with the specific intention of determining whether some specific glyphs, in some cases, might be punctuation.

In 2011, researchers Sravana Reddy and Kevin Knight published a paper on the Voynich manuscript, in which they concluded inter alia that there was no sign of punctuation in the manuscript. Their criterion was that in most Indo-European languages a punctuation mark occurs at the end of a word and can be removed to leave another word. For example, from the word "why?," we can remove "?" and leave the word "why."

In our own testing of this hypothesis, we found several phenomena which looked like punctuation. Our methodology was as follows:

- to set aside the "words" of one or two glyphs, assuming them to have intrinsic meaning and neither to be nor to contain punctuation; in other words, to be irreducible building blocks of the Voynich manuscript
- to identify and list the 250 most frequent distinct "words" of three glyphs or more

- for each of these "words," to remove the initial and final glyph and see what remained
- to count occurrences of the residual string as a whole "word" in the manuscript.

- to select specific glyphs or glyph strings which often appeared at the beginning or end of Voynich "words"
  - for each such glyph or string, to identify the "words" which began or ended with that glyph or string, and to list the 100 most frequent distinct "words"[21]
  - for each of these "words," to remove the glyph or string in question and see what remained
  - to count occurrences of the residual string as a whole "word" in the manuscript

Table 5.11. Voynich manuscript, v101 transliteration, "words" of three or more glyphs: effect of removing selected initial or final glyphs or glyph strings; and comparison with modern English

string	v101	position	number of words containing string	number of words tested	words with string removed				ex
					count	percent	word	v101	wor cou
Voynich manuscript any									
any		initial & final	32,956	200	166	83%	ctoy	1oe	315
any		final	32,956	200	159	80%	oy	oe	710
o	9	final	3,605	100	75	75%	ch2c9	2c9	209
r	y	final	1,085	100	71	71%	ctor	1oy	229
o9	89	final	1,015	100	53	53%	ch2c89	2c89	219
y	e	final	993	100	68	68%	ctoy	1oe	315
aiin	m	final	588	100	32	32%	o aiin	8om	16
s	s	final	508	100	71	71%	ctcos	1cos	32
any		initial	32,956	250	244	98%	oy	oe	710
o	o	initial	2,014	100	100	100%	ol aiin	oham	196

ᑕ	1	initial	1,184	100	80	80%	ᑕᐝᔭ	1oe	402
ᖶᐝ	4o	initial	855	100	99	99%	ᖶᐝᖴᐧᐦᐁ	4ohan	264
ᔮ	9	initial	634	100	99	99%	ᔮᖴᑕᔮ	9hC9	41
ᔕ	8	initial	573	100	94	94%	ᔕᐧᐦᐁ	8am	739
ᑦᑐ	2	initial	497	100	74	74%	ᑦᑐᔭ	2oe	141
ᖴ	k	initial	406	100	85	85%	ᖴᐧᐦᐁ	kam	46
ᔭ	e	initial	385	100	100	100%	ᔭᑕᑕᔕᔮ	e1c89	74
ᖶ	h	initial	378	100	83	83%	ᖶᐧᐦᐁ	ham	96
ᐧ	a	initial	273	100	65	65%	ᐧᐣ	ay	489
ᔕ	8	penultimate	1,015	100	92	92%	ᑕᑕᔕᔮ	1c89	360
ᑕ	c	penultimate	513	100	95	95%	ᑕᑕᔮ	1c9	381

modern English (Brown University corpus)

any letter	final	46,882	200	39	20%	with	7,28
s	final	9,331	100	70	70%	its	1,85
e	final	5,709	100	10	10%	one	3,35
d	final	5,673	100	17	17%	and	28,87
ed	final	4,567	100	50	50%	called	401
n	final	3,733	100	16	16%	been	2,47
t	final	2,776	100	16	16%	not	4,86
es	final	2,498	100	16	16%	does	573
any letter	initial	46,882	200	36	18%	that	10,78
s	initial	5,182	100	18	18%	she	2,98
c	initial	4,159	100	13	13%	can	2,03
p	initial	3,408	100	20	20%	president	410
d	initial	2,570	100	12	12%	down	895
e	initial	1,818	100	7	7%	every	491

Source: author's analysis. In this table, "word" denotes a string of three or more glyphs, delimited by spaces, uncertain spaces or line breaks; or an American English word appearing in the Brown University corpus.

reducible in 50% of cases or more

reducible in less than 50% of cases

If in such a test, all or most of the residual strings were "words," in our view this would be evidence that the glyph or string that we had removed was some form of punctuation.

We tested this methodology on the Voynich manuscript, using the most-common initial and final glyphs as well as the initial ⊤o {4o} and the terminal ∞9 {89} and c9 {c9}; and as a control, on the Brown University corpus of modern English, using the most-common initial and final letters as well as the final "-es" and "-ed." Our results are summarized in Table 5.11.

What these results told us was that:

- Of the 250 most common "words" of three glyphs or more, 98 percent could lose the initial glyph, 80 percent could lose the final glyph, and 83 percent could lose both the initial and the final glyph, and still leave a Voynich "word."
- An initial o {o}, cc {1}, ⊤o {4o}, 9 {9}, ∞ {8}, ⊓ {k}, ϱ [e], ⊓ {h}, or a {a} could be removed from a "word," and in 73 to 100 percent of such cases the remainder was another "word."
- A final 9 {9}, ϒ {y}, ϱ {e}, or ϩ {s} could be removed from a "word," and in 71 to 82 percent of such cases the remainder was another "word."
- A final ⌇⌇ϑ {m} generally could not be removed from a "word." However (although not shown in Table 5.11), a final ⌇⌇ϑ {m} could be replaced by a final ⌇ϑ {n}, which differs only in the absence of the first quill stroke; and in 72 percent of such cases the result was a "word."
- Some penultimate glyphs could be removed. If the "word" ended in ∞9 {89} or c9 {c9}, and only the penultimate ∞ {8} or c {c} was removed, the result was nearly always a "word."
- Most of these phenomena seemed to have no parallel in modern English. The closest comparison was the corpus of English

> words ending with “s” (nearly always denoting the plural), of which 70 percent could lose the “s” and still leave an English word (generally the singular form).

In English literature, these phenomena would be quite extraordinary. It is as if we could remove the initial and final letters of most of the words of Shelley’s “Ozymandias”:

> I met a traveller from an antique land
> Who said: two vast and trunkless legs of stone
> Stand in the desert; near them, on the sand,
> Half sunk, a shattered visage lies

leaving a sequence something like this:

> I e a ravelle ro an ntiqu an
> h ai: w as n runkles eg of ton
> tan in h eser; ea he, on h an,
> al un, a hattere isag ie

and that the mutilated sequence would still mostly consist of English words.

We were impelled to wonder whether the initial, final, and penultimate glyphs and glyph strings that we had tested, and the initial quill strokes of glyphs such as ⟑ [{m}, and probably other glyphs or glyph strings yet to be identified, were some forms of punctuation, or at least had no intrinsic meaning. We might call them “redundant.”

It seemed possible that in the Voynich manuscript, concepts that we consider as punctuation had been transcribed as glyphs or as quill strokes, but with no sign by which the modern reader could distinguish their special functions.

For the final “redundant” glyphs, we have possible counterparts in modern English punctuation, such as the period, comma, colon, semicolon,

question mark, exclamation mark, closing bracket and closing quotation mark, all of which we expect to find at the ends of words.

We could also consider the possibility that the final "redundant" glyphs represent various forms of the plural, like the English "-s" or the French "-s" and "-x," which if removed will generally leave the singular form of the word.

For the initial "redundant" glyphs, we have far fewer counterparts in modern English punctuation, principally the opening bracket and the opening quotation mark, and no obvious counterparts analogous to the plural ending.

An alternative interpretation is that the initial "redundant" glyphs or strings represent the definite article ("the" in English), or the indefinite article ("a" or "an" in English) but attached to the corresponding noun in a way that is rare in Indo-European languages. There are at least eight modern Indo-European languages which attach the definite article, but always as a suffix:[22] for example, in Romanian "barbat" means "man" and "barbatul" means "the man." To this author's knowledge, the only European language which attaches the definite article as a prefix is Maltese (a Semitic language), in which "the" generally translates to "il-" or "l-," and in some cases assimilates with the following consonant (for example "the fire" translates to "in-nar").

For the penultimate ﻫ {8} and ﻛ {c}, things are not so simple. In English, we have the apostrophe, which is often the penultimate character. It has at least two variants. One is as an abbreviation, in colloquial expressions such as "it's" ("it is"), or "we're" ("we are"), or "doesn't" ("does not"); the apostrophe generally represents a vowel, but not a specific one. The other variant is the possessive case, in both colloquial and formal expressions such as "John's" ("belonging to John"). Other Indo-European languages use similar constructions in colloquial speech: for example, the French "j'ai" ("I have") and the Romanian "n-are" (there is not"). If the Voynich manuscript uses the apostrophe in this way, it may represent the possessive case, or it may be derived from documents written in colloquial forms of the source languages.

Yet another possibility is that the "redundant" glyphs are simply junk, as Capt. Currier suggested, inserted throughout the text for an unknown purpose.

So far, we have looked only at the 250 most frequent “words” of three glyphs or more, and for each glyph or glyph string that we have tested, at most the 100 most frequent “words” that started or ended with that glyph or string. On that basis, of the 32,956 such “words” in the Voynich manuscript, we identified at least 21,192 which could lose an initial glyph or glyph string, and/or a penultimate glyph, and/or a final glyph or glyph string, and still leave another “word.” We could call these “words” reducible. It seemed possible that the process of reduction (the removal of glyphs or glyph strings) would not substantially alter any meaning that these “words” might possess.

We could now consider the following hypothesis:

- that all “words” of three glyphs or more are reducible (by removing either the initial glyph or glyph string, or a penultimate glyph, or the final glyph or glyph string, or some combination of these)

If this hypothesis proved true, it would imply that up to 81 percent of the “words” in the manuscript are reducible. Expressing the same idea in terms of the glyph count: of the 158,947 glyphs in the v101 transliteration, we might find that ultimately, we could remove approximately 66,000, or 41 percent of the entire text, without losing any substantive meaning. That would greatly simplify our search for the meaning of the manuscript, if meaning is present.[23]

Our analysis has suggested that most of the common “words” in the Voynich manuscript are reducible. We do not know to what extent the Voynich “words” will continue to be reducible if we look at the less common ones. A rigorous test of this hypothesis would require checking all the 9,370 distinct “words” in the manuscript. That is a task for another day, or an adventure for those of our readers who are so inclined.

CHAPTER 6

D'IMPERIO

Figure 6.1. Mary E. D'Imperio. *Station Hypo*

It is to Mary E. D'Imperio that we owe the most comprehensive history of the Voynich manuscript. She completed her monograph *The Voynich Manuscript: An Elegant Enigma*[1] in 1978, under the auspices of the National Security Agency where she was then employed. Her book remains today the single most complete record of the origin and characteristics of the manuscript, its glyphs, and its illustrations, as well as incorporating an exhaustive account of the research from 1912 to the late 1970s on the possible meaning or purpose of the manuscript.

In her introduction to *An Elegant Enigma*, she wrote:

> I have unwittingly retraced the steps of all my predecessors, rediscovering their sources, repeating their experiments, growing excited over the same promising leads that excited them, and learning only later that all these things had already been tried and had failed, often several times…. in fact, I little

> suspected, when I was first introduced to the problem of the Voynich manuscript at Brigadier Tiltman's lecture in November 1975, that I would spend all my spare time for the next year on an intellectual and spiritual journey spanning so many centuries and ranging over so many aspects of art, history, philosophy, and philology.[2]

Considering that she was one of the great chroniclers of the Voynich manuscript, we have found only the barest outline of Mary D'Imperio's life. She was born in 1930 in Germantown, Pennsylvania. She graduated from Radcliffe College with a degree in comparative philology and classics, and from the University of Pennsylvania with a degree in structural linguistics.[3] As far as we can determine, she worked for agencies of the US government from 1951 until her retirement. She first began work on the Voynich manuscript in 1975; in November 1976, she hosted a groundbreaking seminar in Washington, DC, titled "New Research on the Voynich Manuscript." As well as *An Elegant Enigma*, Mary D'Imperio wrote at least three articles on technical aspects of the manuscript: one in June 1978,[4] and the others (both undated, and one of them classified "Top Secret Umbra") probably shortly thereafter.[5,6] After *An Elegant Enigma*, as far as we can determine, she never returned to the Voynich manuscript.

In 2002 and 2009, in response to FOIA requests, the National Security Agency declassified two of her technical articles, although with redactions; in 2017 they approved the release of a third article (again with substantial redactions), and the proceedings of the 1976 seminar (without obvious redactions).

Mary D'Imperio passed away in 2020 in Springfield, Virginia.

For a full account of the history and mystery of the Voynich manuscript, at least as it was known up to the late 1970s, the interested reader can do no better than to read *An Elegant Enigma*, which is available online from the portal of the National Security Agency.[7] In this chapter, we shall endeavour simply to recapitulate Mary D'Imperio's work of scholarship, and to highlight the subsequent advances in Voynich research.

Mary D'Imperio's work on the Voynich manuscript was inevitably a product of its time. When she embarked on her research, the internet did not exist. As she acknowledged, she was compelled to repeat the errors that

others had made, and to travel false trails that others had followed. Personal computers were in their infancy. She and her colleagues must have had access to computers; they were undoubtedly what used to be called mainframes, occupying large, air-conditioned rooms, subject to programming and scheduling requests and with competing priorities. For any kind of statistical analysis of the manuscript, she had to rely on samples of the text, or on the work of colleagues (notably Prescott Currier).

In those days, image scanning was also an infant technology. We know from D'Imperio's writings that her only exposure to the original Voynich glyphs and illustrations was through black-and-white photocopies, or "photostats" as they were called in those dear departed days.

D'Imperio devoted considerable effort to study of the supposed link between the manuscript and Roger Bacon. She could not have known that the Voynich parchment would eventually be submitted to radiocarbon technology and that the samples would be dated, with up to 92 percent probability, to periods ranging between 1308 and 1458. Therefore, she could not have known that Bacon, who lived in the thirteenth century, would be excluded as the author of the manuscript, or at least as its producer or as one of its scribes.

IMPRESSIONS

In *An Elegant Enigma*, D'Imperio set out the following areas of her research:

- the known facts relating to the manuscript (some of which, strictly speaking, were claims made by Voynich and others about its origin and ownership)
- a survey of general propositions advanced at various times, as to the authorship of the manuscript, its purpose, its geographical origin, and its content and underlying languages (if any)
- a meticulous description of the illustrations, in each of their thematic sections, accompanied in an index by many hand-

drawn representations of key elements of the illustrations, as she perceived them (an example is shown in Figure 6.2)

- an equally exhaustive examination of the glyphs, their construction and style, and their relationship (if any) to the languages and scripts of other medieval documents
- two chapters relating to previous authors' claims to have deciphered parts of the manuscript, and to other substantive research and analysis during the period from 1912 to 1976
- four chapters intended to provide a context for the manuscript, relating respectively to Roger Bacon, medieval cosmology and iconography, artificial and secret languages, and medieval herbal texts.

Finally, she proposed avenues for further research, including palaeographic analysis, more investigation of the history of the manuscript, and development of a complete machine-readable transliteration. In each of these three specific areas, progress has been made, as we shall recount later in this chapter.

Figure 6.2. Mary D'Imperio's hand-drawn sketch from a photocopy of page f023r, and the corresponding detail from the scanned image. *Mary D'Imperio and Beinecke Rare Book and Manuscript Library*

D'Imperio's work is eminently readable today, notwithstanding the forty-four years that have elapsed since it was written. The information that

she assembled is authoritatively presented and comprehensively documented. In *An Elegant Enigma*, she seemed to keep an open mind and did not seem to advance a preferred hypothesis or interpretation of the Voynich manuscript. (In her technical articles, she did present certain of her own conclusions, to which we shall return). To this author's mind, in the context of a search for meaning in the manuscript, the most interesting passages of *An Elegant Enigma* relate to D'Imperio's personal and subjective impressions of the manuscript.

In general terms, D'Imperio was in no doubt that the manuscript was worth investigation:

> The Voynich manuscript appears to be unusual in that it combines in one book at least four different medieval disciplines, apparently with some attempt to integrate them into a single system.
>
> No one, to my knowledge. has so far discovered anything else at all like it. It strikes the viewer as a very strong and definite statement, completely independent of any known style or doctrine.
>
> If the manuscript is a compilation, however "deranged" or idiosyncratic, drawn from earlier magical, alchemical, or medical works, it has at least as much intrinsic interest and "scientific" import for the history of Western thought as do other similar manuscripts which are readable, and concern only one topic (i.e., they are either astrological, or alchemical or medical).
>
> … even if the text is totally meaningless (a possibility that seems to me highly unlikely), a decipherment of the text in some manner permitting an understanding of the code, cipher, or other concealment system employed should be of great interest for the history of cryptology, and perhaps also for the study of alphabets and writing systems.[8]

On the visual impact of the document, even though she was working only with photocopies, D'Imperio shares with us the sense of wonder and bewilderment that she felt, and which today's reader may also feel. She also

expressed her respect for the aesthetic qualities of the work; it did not seem to her the "ugly duckling" that Voynich called it:

> The impression made upon the modern viewer first coming upon a photocopy of the manuscript … is one of extreme oddity, quaintness, and foreignness—one might almost say unearthliness.
>
> The drawings and text … form a consistent whole, the product of one school or group of closely related persons if not of a single person.
>
> Many of the plant folios and some cosmological designs … present a stalwart, bold felicity of composition that is almost architectonic in its quality, and (to me) quite pleasing … [it is] craftsmanship rather than art.
>
> [there are] rules and relationships, a definite structure with its own "logic," however erratic and bizarre … [we see] intricate compound forms in the script and its matter of fact, rather austere style … there is a persistent tectonic element of style in the drawings, emphasizing threedimensional forms, symmetry, and connectedness of parts.[9]

When it came to the glyphs, D'Imperio was greatly impressed (as was this author) by the uniqueness, fluency, and harmony of their construction:

> It has a deceptively flowing, rhythmic quality that suggests long practice and familiarity on the part of the scribe or scribes…. a look of ease and natural flow … an intricate but structurally logical system of ligaturing or compounding of simple forms to build up more complex outlines…. a tour de force of artistry and ingenuity.[10]

She went on to catalog and to describe some of the many previous attempts at transliteration and interpretation of the glyphs, including her own observations on their frequencies, positioning, and construction. In general, she seemed to consider the extensions and embellishments of

certain glyphs as ornamentation rather than significative of a specific meaning.

THE D'IMPERIO CRITERIA

Mary D'Imperio did not, in *An Elegant Enigma*, advance any personal view of what the glyphs meant (if they meant anything). However, she did enumerate fifteen phenomena (and some of these, although she did not say which ones, were her personal choices) which she felt any analyst of the Voynich glyphs would have to consider. Since the interpretation of the glyphs is central to the present project, D'Imperio's fifteen phenomena follow below (the numbering is hers, and the comments are this author's):

> "*(1) The basic alphabet of frequently occurring symbols is small (as few as fifteen ... and probably no more than twenty-five)*." [We recall from Chapter 3 that the v101 alphabet has a basic set of seventy-one glyphs, although on the v101 definitions, the twenty most frequent glyphs account for 94.7 percent of all the glyphs in the manuscript.]
>
> "*(2) The basic forms are compounded or ligatured to create a large variety of complex symbols*." [We recall from Chapter 3 that in the v101 alphabet, there are eighteen "gallows glyphs," including "pedestal glyphs" such as {H}, that seem to be built from more basic forms; and that there are at least ten glyphs, such as {m}, that could be reconstrued to be sequences of still-more-basic glyphs.]
>
> "*(3) The symbols are grouped into "words" separated by spaces (although some researchers have expressed doubts about the consistency of these spaces)*." [Again, from Chapter 3: the v101 alphabet recognizes the "space" and "uncertain space"; but to this author's mind, there is a continuous spectrum of spacing between "no space," "uncertain space" and "space." It would take an exercise in optical character recognition to quantify this spectrum.]

"*(4) The number of different words seems surprisingly limited*." [She was probably referring to the Currier transliteration of 1976; in 2011 Reddy and Knight calculated, working from that transliteration, that the Voynich manuscript contained 37,919 "words" of which there were 8,114 distinct "words."[11] By way of comparison: Shakespeare's *A Midsummer Night's Dream* has 16,268 words and 2,931 distinct words.[12] So the vocabulary of the Voynich manuscript is not as obviously limited as D'Imperio perceived.]

"*(5) The "words" are short, averaging around four or five symbols in length*." [But so are they in *A Midsummer Night's Dream*, where the average word length is 4.04 letters.][13]

"*(6) The same "word" is frequently repeated two, three, or more times in immediate succession*." [Here we could challenge the term "frequently"; in Chapter 3 we quoted an estimate that there are about 250 instances of a repeated word, and about 20 instances of a triplet. The literary term is "diatrope." Prose examples abound in classical and popular culture: "Verily, verily, I say unto thee" (Jesus of Nazareth), "Romeo, Romeo, wherefore art thou Romeo" (William Shakespeare), "It is a far, far better thing that I do" (Charles Dickens), "I will build a great, great wall" (Donald Trump). In poetry or song, we would not think twice about repetitions: "Good, good, good, good vibrations" (the Beach Boys), "It's over, it's over, it's over" (Roy Orbison); "Boys Boys Boys" (Lady Gaga).]

"*(7) Many "words" differ from each other by only one or two symbols, and such "words" often occur in immediate succession*." [To evaluate this comment, we would need to understand whether the "words" are unusually long. In English prose, we would not raise an eyebrow over sequences of short but similar words, such as "it is in" or "no not now." In English-language poetry, such a sequence of longer words might well be a rhyme: "Lovely Rita, meter maid" (the Beatles), "Helter Skelter" (the Beatles), "Woolly Bully" (Sam the Sham and the Pharaohs).

"*(8) Certain symbols occur characteristically at the beginnings, middles, and ends of "words," and in certain preferred sequences*." [This is probably true of many European languages. For example, in Italian, Portuguese, and Spanish, the letters "a," "e," "i," and "o" are ubiquitous word endings, signifying gender and number. In most European languages, "q" can only be followed by "u" and can never occur at the end of a word.]

"*(9) Certain symbols appear very rarely, and only on certain pages, indicating some special function or meaning*." [This is true of the symbols "$," "£," "#," "%" and several others on this author's English-language keyboard.]

"*(10) There are very few doublets (repetition of the same letter twice in succession)*." [This depends on how a glyph is defined: for example, the glyph ⟨ɪɪƞ⟩ {m} occurs 4,112 times in the v101 transliteration, and it could be argued that it contains at least two instances of the glyph ⟨ɪ⟩ {i}.]

"*(11) Very few symbols occur singly (as one-letter "words") in running text; these are primarily* *and* ." [By this author's count, the glyphs that occur most frequently in isolation are ⟨ƨ⟩ {s} 449 times, ⟨ϥ⟩ {9} 428 times, ⟨ɾ⟩ {e} 285 times, ⟨ʏ⟩ {y} 284 times, and ⟨o⟩ {o} 264 times. In English, the only single-letter words are "a," "I" and "o." In *A Midsummer Night's Dream* (which has 16,268 words), "I" occurs 442 times, "a" occurs 264 times, and "o" occurs 58 times.[14] The isolated Voynich glyphs do not seem greatly out of line with Shakespeare.]

"*(12) Prefix-like elements are tacked in front of certain "words" that also occur commonly without them: such prefixed elements are* , , *and* ." [This is true of most European languages; in English, common examples are "re-," "un-" and "dis-."]

"*(13) The symbol* *occurs almost invariably followed by* *and joined to it by an extension of the crossbar on the* ." [This is a very valid observation, and in this author's view,

probably indicates that the shape ⟙ {4o} is a single glyph and not two.]

D'Imperio's phenomena (14) and (15) relate to the frequency of certain glyphs as the initial glyphs of the text in the herbal section, and as the initial glyphs of labels in the herbal, cosmic and other sections. Her observations were based on her expectations that these sections should conform to patterns in medieval herbal and other texts; but in this author's view, there is no need for such expectations.

In summary, in this author's view, D'Imperio's observations on the glyphs were largely accurate and valid but did not exclude the existence of an underlying European language or languages with some form of mapping between glyphs and letters.

PTAH

In two of her three published technical articles on the Voynich manuscript, Mary D'Imperio reported on the statistical clustering of the glyphs on certain Voynich pages but did not express a view on the nature of the glyphs.

In her third article,[15] for the first and only time (at least in today's public domain) she drew professional conclusions.

This article is undated but probably was written in the fall of 1978, since it refers to her earlier article published in the summer of 1978, but predates *An Elegant Enigma*, which she mentioned as "my forthcoming monograph." It relates to her application of a computer program called PTAH (not an abbreviation, but named whimsically by a fellow programmer, after the Egyptian god of craftsmen and architects). This article, until it was approved for release in 2009, was classified "Top Secret Umbra."

Why Umbra?:

> … at least since [the] middle of the 1950s—the intelligence community used the UMBRA code word to inform the reader of a certain report that the original source for the intelligence was of the most sensitive category.[16]

D'Imperio described the flavor of the PTAH program as follows:

> Let us imagine an animal that can exhibit five major kinds of activities … He can eat. sleep, hunt for food, sit still, and scratch for fleas. By counting a long sequence of actions in the animal's life. we can arrive at an idea of which sets of actions he is likely to do, in which order…. like an automatic washer going through its cycle…. Our way of looking at the "letters" of the Voynich text … is similar to our view of the mythical five-state animal.[17]

The way that D'Imperio applied the PTAH program to the Voynich manuscript was, in essence, to determine whether the glyphs fell into sequences that approximated to those in natural languages. The idea (if this author understood it correctly) was that in natural languages, certain letters can be followed by some letters but not others. For example, in English a "c" can be followed by any vowel, and by several consonants such as "l," "t," or another "c," but (we suppose) cannot be followed by an "f." D'Imperio was looking to see whether the Voynich glyphs obeyed some such rules.

Working first with the English language, she reported as follows:

> The diagram for Voynich symbols shows only eight significant transitions. while that for English letter[s] shows thirteen.[18]

In plain language, she was saying that the Voynich glyphs did not behave like letters in English.

She then tested her Voynich sample against a text in Latin, with 4,700 letters, extracted from *Magia Naturalis*, by Giovanni Battista Porta, and dated 1644, which she described as "a work concerned with materia

medica, medical spells, natural ‘wonders,’ and such matters which seem related to the apparent content of the Voynich manuscript as evidenced by the drawings.” She reported that:

> … the Latin diagram still has a lot more arrows and a much more intricate set of interconnections than that for Voynich symbols (eleven arrows as compared to eight).[19]

Otherwise expressed: the Voynich glyphs did not behave like Latin letters either.

The second part of D’Imperio’s study related not to individual glyphs but to the behavior of whole “words” or glyph strings (with the reservation that it was not always clear where one glyph string ended, and another began). Here, her conclusions were in essence as follows:

> It is strikingly clear that there is a positional structure of words within lines in the Voynich text and that certain sets of words … are most likely to follow or precede certain other sets of words … Currier has pointed out these two features of the text … Our analysis clearly supports both his view of the lines as functional entities, and his finding that words with certain endings were more likely to be followed by words with certain beginnings within a line. This is strange behavior indeed for any running plain text, unless it represents lists of parallel phrases (incantations? instructions? recipes?) in highly stereotyped form.[20]

In the third part of her study, D’Imperio examined the parts of Voynich “words.” She wanted to see if she could build Voynich “words” from a basic set of simpler elements. She explained the idea as follows:

> … let us imagine that we were “parsing” the English phrase “NOW/IS/THE/TIME/” using [the] strings NO, IS, /, THE and ME: the result would be:
>
> NO W / IS / THE / TI ME /

> with ten product strings, two of which (W and TI) are leftovers.[21]

Maybe an analogy from the construction industry would be more accessible. Suppose we were to demolish a building. At the end of the job, we would have five piles of recyclable materials, say bricks, wooden planks, concrete blocks, reinforcing bar and glass panels: plus, two piles of miscellaneous scrap.

What she was about to ask of PTAH was: could we deconstruct the Voynich building in this way? Or maybe a better question for the Egyptian god of architects and engineers: is the Voynich building made of basic blocks and bricks, or is it made from prefab panels?

Cutting to the chase, she reported that:

> I find these analyses even more convincing in confirming the highly regular positional structure of elements in Voynich text "words." In addition, these results suggest that the meaningful elements are not words as wholes, or single letters, but larger, variable-length sequences of symbols.[22]

This was a profound discovery. D'Imperio was saying that the ingredients of the Voynich manuscript were not the individual glyphs as conventionally understood, but glyph pairs, glyph triplets, maybe glyph quadruplets and more, or mixtures of these. Indeed, she would later report in *An Elegant Enigma*, without quite saying so, that the glyph "pair" ⟙o was in effect a single glyph, or at least a single building block of the manuscript.

At the end of her three-part study, D'Imperio expressed her conclusions in no uncertain terms:

> The plain text directly underlying the Voynich text is probably not a natural language represented by an alphabet of single letters like the English alphabet.
>
> … the encryption or concealment system in the Voynich text probably is not any form of simple substitution on an alphabet

of single letters like the English alphabet.

> The Voynich text probably does not represent a natural language written in an "impressionistic" way … nor can its characteristics be explained by hypothesizing many variant spellings of the same words in an alphabetic writing system (cf. older forms of English).
>
> If the Voynich text conceals a code, it is not very like … a code involving a partially inflected Romance language comparable in some ways to and descended from Latin …
>
> … it is hard to imagine any directly underlying natural language plain text whose characteristics can explain the phenomena adequately.[23]

In short, D'Imperio had her own repeated word: Bang Bang. She shot down all previous research that had sought to look for English, Latin or any natural languages as substrates to the Voynich manuscript.

AFTERMATH

Inexplicably, in *An Elegant Enigma* D'Imperio did not repeat any of her conclusions from the PTAH research. Perhaps her monograph had already gone to the press (it is dated 1978). Perhaps she had second thoughts about PTAH. Perhaps the code word UMBRA on her technical article had something to do with it. In the monograph, she was studiously neutral toward all previous research. Even her fifteen criteria for further analysis of the Voynich glyphs, as we observed above, did not exclude the search for natural languages. After her article on PTAH was classified as secret, thirty-one years would pass before anyone outside the National Security Agency would know of her work. In some ways the story of *An Elegant Enigma* parallels that of D. B. Cooper and the Norjak case: from the hijacking of Flight 305 in 1971, thirty-six years would elapse before the FBI opened their files on the most famous unsolved case in their history.

In any event, this author has kept an open mind with regard to D'Imperio's conclusions, at least those relating to PTAH, and has explored

many avenues of research relating to underlying natural languages. We shall travel some of those avenues in later chapters.

POSTSCRIPT: BUILDING BLOCKS

For what it is worth, this author has taken seriously Mary D'Imperio's conclusions about groups of glyphs as building blocks of the Voynich manuscript, and has counted the occurrences of glyph pairs, glyph triplets, glyph quadruplets and so on, up to glyph 12-strings.

In Table 6.1 we have listed the thirty most common glyph pairs, triplets, and quadruplets in the manuscript. It is no surprise that among the glyph pairs, the ubiquitous glyph ○ {o} appears in the first, second, third and sixth positions, preceding or following another common glyph (and maybe, as D'Imperio implied or suspected, integrated with that glyph in some fashion whereby only the glyph pair has meaning, and the individual glyphs have none).

We have been assuming all along that the scribes started with documents in natural European languages and transliterated these documents into the Voynich glyphs. Let us suppose, as Mary D'Imperio has encouraged us to believe, that glyph pairs, triplets or longer strings, and not individual glyphs, were the building blocks of the Voynich manuscript. In this case, the question arises: what was the process of transliteration from letters to glyphs?

Firstly, we might hypothesize that a single European letter (in Latin, Cyrillic, or some other script) was mapped into a glyph pair; and that a space in the source document was mapped into a space in the Voynich manuscript.

In this case, the Voynich "words" would be twice as long as the European words. Moreover, every Voynich "word" would contain an even number of glyphs, which is plainly not the case. So immediately we must modify our assumption, to the effect that some letters were mapped to glyph pairs, and some letters were mapped to single glyphs.

We calculated in Chapter 3 that the v101 transliteration of the Voynich manuscript contained 40,708 words, made up of 158,947 glyphs. In this

transliteration, the average length of a Voynich "word" was 3.91 glyphs.

Many researchers have remarked that the words in the Voynich manuscript are extraordinarily short, by the standards of European language corpora. Just how short, we shall now find out. We made a comparison with the average word length in modern European languages, plus Hebrew and selected medieval documents. This comparison is summarized in Table 6.2.

This compilation confirms that the Voynich manuscript has, on the average, much shorter words than any modern European language. The modern language that comes closest to matching it is Hebrew, with an average word length of 4.40 characters (as a consequence of omitting all short vowels). However, some medieval documents that we have identified have similar word lengths: for example, Shakespeare's *Macbeth, A Merchant of Venice*, and *A Midsummer Night's Dream* with averages of about four letters, and the epic Old French romance poem *Partonopeus de Blois* with an average of 3.97 letters.

Table 6.1. Voynich manuscript, v101 transliteration: the twenty-five most frequent glyph pairs, triplets and quadruplets (excluding spaces and uncertain spaces).

pair	v101	G2 count
ol	oh	5,863
oe	oe	5,406
to	4o	5,250
89	89	4,886
cc	1c	4,547
ok	ok	3,905
am	am	3,794
8a	8a	3,288
ay	ay	3,235
ae	ae	3,110
ha	ha	2,933
c8	c8	2,908
co	co	2,889
c9	c9	2,598

ᴄᴄo	1o	2,588
ꟻc	hc	2,454
oʏ	oy	2,441
o8	o8	1,872
89	79	1,748
aιɔ	an	1,714
cc	cc	1,660
ᴄ²c	2c	1,628
ꟻa	ka	1,569
ꟻc	kc	1,389
ꟻcc	hC	1,329
triplet	v101	G3 count
ꜝoꟻ	4oh	3,013
c89	c89	2,287
oꟻa	oha	1,844
oꟻc	ohc	1,547
ꜝoꟻ	4ok	1,139
8aιɔ	8am	1,131
oꟻa	oka	1,116
ᴄᴄ8	1c8	1,101
ᴄᴄo	1co	1,034
c89	c79	1,025
oꟻc	okc	974
ᴄᴄ9	1c9	870
oꟻcc	ohC	833
coɤ	coe	815
ᴄoɤ	1oe	777
ꟻaιɔ	ham	750
ꟻcc	hcc	685
8aʏ	8ay	665

hae	hae	637
han	han	637
co8	co8	615
8ae	8ae	610
hco	hco	560
hay	hay	548
ok1	ok1	542
quadruplet	v101	G4 count
4oha	4oha	952
1c89	1c89	833
4ohc	4ohc	792
4ohC	4ohC	539
ohae	ohae	469
oham	oham	456
ohan	ohan	419
ohcc	ohcc	387
ohco	ohco	363
4oka	4oka	315
ohay	ohay	314
okae	okae	291
1coe	1coe	291
ohc8	ohc8	290
4okc	4okc	256
okay	okay	250
okam	okam	249
4oh1	4oh1	245
4ok1	4ok1	201
ohc9	ohc9	181
4oho	4oho	181
1co8	1co8	171

okan	okan	169
ohc7	ohc7	161
4oh9	4oh9	143

Source: author's analysis based on counts from v101 transliteration in Microsoft Word

However, even if the Voynich manuscript was derived from documents in the style of Old French poetry or Shakespeare's plays, it would not have been possible to map single letters to glyph pairs, without greatly increasing the lengths of the glyph strings. Such a mapping would require a source document with an average word length of about two characters; and no such document is known to exist.

In order to understand Mary D'Imperio's observations on "parsing" an English phrase into its building blocks, we have to imagine that digraphs (letter pairs) in the source document were mapped to glyph pairs in the Voynich manuscript; and that there was no unique mapping of the first letter to the first glyph, or of the second letter to the second glyph. As a purely hypothetical example, the Voynich scribes might have used their two-glyph building blocks (as illustrated in Table 6.1) to transliterate "tomorrow" as follows:

to = 4o mo = 89 rr = a ii ow = ch

Table 6.2. Average lengths of words in selected corpora of modern European languages and Latin; and in selected medieval documents

language	characters	words	av. word length
modern language corpora			
Ukrainian	2,655,232	360,344	7.37
Latvian	2,005,585	274,484	7.31
Belarussian	3,006,424	415,358	7.24
Latin	2,059,595	285,103	7.22
Finnish	2,058,115	288,202	7.14
Croatian	2,471,874	365,664	6.76
Turkish	1,947,735	294,656	6.61

Macedonian	1,996,246	302,511	6.60
German	8,953,540	1,368,984	6.54
Polish	2,020,506	309,880	6.52
Russian	1,351,370	210,844	6.41
Estonian	2,051,977	321,323	6.39
Albanian	3,470,343	562,365	6.17
Greek	1,442,382	235,614	6.12
Italian	2,079,288	344,048	6.04
Czech	1,885,069	315,462	5.98
Swedish	2,036,475	351,676	5.79
French	2,182,037	377,959	5.77
Dutch	2,055,683	356,504	5.77
Icelandic	1,857,253	323,562	5.74
Galician	2,062,295	360,712	5.72
Danish	2,073,451	370,653	5.59
Spanish	2,129,690	381,084	5.59
Norwegian	2,037,670	366,595	5.56
Irish	2,079,887	374,555	5.55
English	2,101,261	380,703	5.52
Catalan	2,060,555	377,351	5.46
Welsh	2,044,201	379,806	5.38
medieval corpora and documents			
Latin, 1313-14			
Monarchia, Books I-III	111,691	19,347	5.77
Bohemian, c.1300-14			
Dalimilova kronika	131,200	25,460	5.15
Early New High German, c.1401			
Der Ackerman aus Bohmen	50,760	10,232	4.96
Italian, 1308-53			
Il Decamerone (1348-53)	1,257,963	255,526	4.92
La Divina Commedia (1308-21)	444,962	100,521	4.43

Albanian, 1555			
Meshari (Missale)	326,356	75,699	4.31
Middle French, c.1457-70			
La Farce de Maistre Pathelin	39,436	9,501	4.15
Arabic, 15th century			
CLAUDIA corpus	~220 million	53,365,460	4.12
Early Modern English, 1595-1601			
Hamlet	120,866	29,770	4.06
The Merchant of Venice	84,213	21,159	3.98
Voynich manuscript, >1438?			
v101 transliteration	158,963	40,702	3.91
v101 with separated words	158,963	*43,951*	*3.62*
Galician-Portuguese, c.1220-1300			
Cantigos d'Amigo	509,880	135,574	3.76
Middle English, c.1340-1500			
Auchinleck manuscript (c.1340)	217,020	59,323	3.66
Ashmole 61, Parts 1-27 (c.1500)	209,710	53,707	3.90
Persian (c.1048-1131)			
Omar Khayyam corpus	30,322	9,088	3.34

Source: (modern languages) Stefan Trost Media; (early modern English) Wolfram Alpha; (Arabic) Dr Jiří Milička; (separated word count) Massimilano Zattera; (others) author's analysis.

Romance group

Germanic group

Slavic group

And we would search in vain for single glyphs that corresponded respectively to the Latin letters "m," "o," r, "t," and "w."

If we are to investigate whether glyph pairs or triplets, or more, could have been the building blocks of the Voynich manuscript, we shall have to look not only at letter frequencies in our target languages, but also at the

frequencies of digraphs, trigraphs and more. This we shall do in another chapter.

POSTSCRIPT: AFTER D'IMPERIO

In the forty-four years since Mary D'Imperio ended her work on the Voynich manuscript (at least, that part of her work that has entered the public domain), some advances have been made in our knowledge and understanding of the manuscript. Specifically:

- Radiocarbon analysis established date ranges, with up to 92 percent probability, for five samples of the parchment; these ranges spanned the period from 1308 to 1458.
- René Zandbergen tracked down further correspondence from Georgius Barschius to Athanasius Kircher, which confirmed that Barschius had indeed sent handmade copies of certain undeciphered writing to Kircher (though we cannot determine that any of this correspondence related to the Voynich manuscript).
- Glen Claston and others developed complete computer-readable transliterations of all the glyphs in the manuscript (although, as we observed earlier, machine readability places a barrier between the researcher and the original document, and also obscures the assumptions that were necessarily made in mapping the glyphs to keyboard characters).

In more general terms, the modern reader now has access to high-resolution scanned images of every page of the document, as well as to cheap computing power such as Mary D'Imperio could not have dreamed of. We will attempt to put these advances to good use in the chapters that follow.

CHAPTER 7

MARKOV

Figure 7.1. Andrei Andreyevich Markov in 1886. *Public domain*

The reader might well enquire: who is Markov and what is his (or her) connection with the Voynich manuscript?

Let us then introduce the Russian mathematician Andrei Andreyevich Markov (1856–1922). He was a pioneer in probability theory, and in particular in the mathematics of processes with probability distributions. His work led to the concepts now known variously as Markov processes, Markov chains or Markov decisions.

A Markov process may be described in simple terms as a mathematical model of a sequence of events, each of which has a set of possible outcomes. The outcomes of each event have probabilities which depend upon the outcome of the previous event. The link between outcomes of consecutive events is the concept which gives rise to the alternative term "Markov chain."

There are some intriguing, but no doubt coincidental, intersections between the lives of Wilfrid Voynich and Andrei Markov. Like Voynich,

Markov was born a subject of the Russian Empire. Like Voynich, Markov was politically active and supported liberal movements in Russia; unlike Voynich, he seems never to have fallen foul of the authorities. Both were scientists: Markov an applied mathematician, Voynich an analytical chemist. In different ways, both were involved in the study of literature. Both had concepts named after them, which survive to this day. The two were contemporaries, although one supposes that they never met.

It was in 1906 that Markov first published his work on the processes that would later bear his name. In 1913, he applied his ideas to a study[1] of the novel *Eugene Onegin* by Alexander Pushkin. He was interested in the predictability of vowels and consonants in literary works. In his model, the "event" was the occurrence of a character in (presumably) Cyrillic script; for each event there were two possible "states": a vowel, or a consonant. With some probability (to be determined), a vowel could be followed by another vowel, or by a consonant; a consonant could be followed by another consonant, or by a vowel. These four alternatives gave rise to a 2 × 2 matrix of probabilities.

Professors Charles Grinstead and Laurie Snell, in their book *Introduction to Probability*, quote the Soviet historian Adolph-Andrei Pavlovich Yushkevich on Markov:

> Markov arrived at his chains starting from the internal needs of probability theory, and he never wrote about their applications to physical science. For him the only real examples of the chains were literary texts, where the two states denoted the vowels and consonants.[2]

Grinstead and Snell described Markov's experiment as follows:

> Markov chose a sequence of 20,000 letters from Pushkin's *Eugene Onegin* to see if this sequence can be approximately considered a simple chain. He obtained the Markov chain with transition matrix:

	vowel	consonant
vowel	.128	.872

consonant	.663	.337

The fixed vector for this chain is (.432, .568), indicating that we should expect about 43.2 percent vowels and 56.8 percent consonants in the novel, which was borne out by the actual count.

Here we may begin to see why the work of Markov is relevant to the Voynich manuscript. Like Markov's view of *Eugene Onegin*, the manuscript may be viewed as a sequence of events.

Taking Glen Claston's v101 transliteration as our source document, we can define an event as the occurrence, or if we prefer, the observation of a glyph, or a space, or a line break. In v101, for any given event, we can define seventy-four possible states: the seventy-one glyphs of the basic set, the space, the "uncertain space," a "rare glyph" (which we can consider as a single state), and a line break. (The transliteration does not have a symbol for page breaks, although it could be adapted to do so.)

Each glyph has a set of probabilities regarding what follows it: the same glyph, or another glyph, or a space, or an "uncertain space," or the end of a line. From Currier's work as reported in the 1976 seminar, we know that some of these probabilities are large. From our own analysis, in chapter 3 we gave the example of the glyph ⟙ {4}, which has a 96.7 percent probability of being followed by the glyph ⍥ {o}.

To give a flavor of how pervasive these patterns are, Table 7.1 summarizes the Markov matrix for the nineteen most common glyphs, plus the space and "uncertain space," in the Voynich manuscript. For each of these "states," the cells to the right give the probabilities of the following "state." Probabilities over 20 percent are highlighted.

Undoubtedly some of these patterns are examples of what Prescott Currier reported in the 1976 Voynich seminar. Among these patterns, in addition to the extraordinary case of ⟙⍥ {4o} which we mentioned above, we are struck immediately by the following:

- The glyphs ⍜ {9}, ⍧ {e}, ⍲ {y}, ⍣ {m}, ⍤ {s}, and ⍢ (n) are likely, with probabilities ranging from 43 percent to 88 percent,

to be followed by a space. Put another way: they are likely, even very likely, to be the final glyphs of a "word." (These percentages in fact are understated, since they do not include instances where the glyph in question is followed by an "uncertain space" or a line break).

- Indeed, even if we did not know this from the glyph counts, even the shapes of these glyphs, with their backward sweep of the final quill stroke, suggest to us that these glyphs are the ends of "words."
- The space is quite often followed by the glyph o {o}. We could equally well say: o {o} is quite likely to be preceded by a space. In other words, o {o} is a common initial glyph of a Voynich "word." However, as we shall see later, its position is more usually internal to a "word."
- The glyphs c {c}, cr {1} and c2r {2} are often followed by o {o}.
- The glyphs 8 {8}, cc {C} and 8 {7} are often followed by 9 {9}. This pattern is reinforced by the possibility that, as we recall from chapter 3, 8 {7} may be just a variant of 8 {8}.
- The glyphs 8 {8}, ⁋ {h}, ⁋ {k} and 8 {7} are often followed by a {a}.

and so on …

Table 7.1. Voynich manuscript, v101 transliteration: a Markov matrix of two-character strings, for the fifteen most common glyphs, plus the space and "uncertain space"

					following glyph/space							
glyph	v101	count	space	space?	o	9	a	c	cr	e	8	
			.	,	o	9	a	c	1	e	8	
space	.	32,475	0%	0%	22%	3%	6%	0%	18%	4%	7%	
space?	,	2,431	0%	0%	8%	4%	24%	2%	11%	6%	10%	1
o	o	24,745	4%	2%	0%	0%	1%	1%	1%	22%	8%	2
9	9	17,309	74%	2%	0%	0%	0%	0%	1%	0%	1%	

ɑ	a	14,296	0%	0%	0%	0%	0%	0%	0%	22%	0%	(
ɔ	c	13,573	1%	0%	21%	19%	3%	12%	1%	0%	21%	:
ƈ	1	11,129	0%	0%	23%	9%	4%	41%	0%	1%	6%	'
ɾ	e	10,518	49%	6%	5%	5%	3%	0%	6%	0%	3%	9
ȣ	8	10,363	4%	1%	4%	47%	32%	1%	3%	1%	0%	(
ʰ	h	9,837	1%	1%	7%	7%	30%	25%	10%	0%	0%	(
ʔ	y	6,635	70%	4%	4%	4%	7%	0%	1%	0%	0%	(
ᶠ	k	6,030	1%	1%	12%	8%	26%	23%	16%	0%	0%	(
ϯ	4	5,469	0%	0%	96%	0%	0%	1%	0%	0%	0%	(
ᥱ	m	4,063	88%	1%	0%	0%	0%	0%	0%	0%	0%	(
ƈ̀	2	3,608	1%	0%	21%	6%	3%	45%	1%	0%	4%	'
ƈ	C	2,800	1%	0%	14%	43%	2%	1%	1%	0%	21%	'
ȣ	7	2,722	2%	1%	1%	64%	24%	1%	1%	0%	0%	(
ƨ	s	2,655	43%	6%	12%	4%	18%	2%	3%	0%	0%	'
ᥱ	n	1,758	88%	0%	0%	0%	0%	0%	0%	0%	0%	(

Source: author's analysis. Matrix excludes line breaks (for which there is no keyboard assignment in the v101 transliteration).

probability 20-40%

probability 40-70%

probability over 70%

The question then arises: are these patterns normal? Or more specifically: do similar patterns exist in natural languages?

We know that in the English language, and indeed in most European languages, including Latin from which the pattern arose, the letter "q" is invariably followed by the letter "u." A native English speaker would probably be hard-pressed to cite another letter whose following letter is generally predictable. For what it is worth, the two most common digrams in English are "th" (3.99 percent) and "he" (3.65 percent).[3] both being components of the word "the" as well as of many others. Perhaps this is one

of the few cases where an English speaker, seeing the first two letters, could guess what the next letter would be.

Let us then look at Latin, for which we have large corpora readily available. Table 7.2 summarizes, for each of the ten most common letters in classical Latin, the frequencies of the digrams beginning with the letter in question.

From this table we can see firstly that in classical Latin, a word break very often follows the letters “s” and “m,” and often follows the letters “e,” “t,” and “o.” Put another way: these letters are frequently the terminal letters of Latin words. Secondly, among digrams made up of two Latin letters, only the digrams “re” and “nt” are unusually frequent. Otherwise, the frequencies are generally low.

In other words, the predictability of two-glyph strings that we have seen in the Voynich manuscript is nearly absent in classical Latin. For any given Latin letter, we do not have much chance of guessing what the following letter will be.

Comparing the Markov matrixes for the Voynich manuscript and for the classical Latin corpus, we can see that the Voynich, as it is usually transliterated, has a structure quite unlike that of Latin. Recalling Mary D’Imperio’s analogy: it looks and feels like a different animal altogether.

Table 7.2. A Markov matrix of classical Latin digrams starting with the ten most frequent Latin letters or a word break

first character	count	second character										
		word break	i,j	e	u,v	a	t	s	r	n	o	
word break	7,494,720	2%	8%	10%	6%	9%	4%	9%	3%	5%	2%	
i, j	4,708,691	12%	1%	2%	7%	6%	13%	13%	2%	13%	5%	
e	4,895,633	21%	1%	0%	1%	2%	10%	10%	17%	10%	1%	
u, v	3,993,819	1%	13%	12%	1%	7%	6%	14%	7%	5%	5%	1
a	3,625,579	16%	1%	9%	5%	0%	14%	5%	7%	9%	0%	1
t	3,485,112	31%	16%	14%	15%	12%	1%	0%	5%	0%	5%	
s	3,193,352	40%	11%	11%	7%	4%	11%	7%	0%	0%	2%	

r	2,641,646	11%	18%	22%	9%	15%	5%	2%	2%	1%	8%
n	2,579,654	12%	13%	12%	7%	7%	20%	5%	0%	1%	9%
o	2,313,475	21%	0%	1%	1%	0%	3%	10%	16%	18%	0%
m	2,292,272	51%	10%	8%	6%	9%	0%	0%	0%	2%	6%

Source: author's analysis based on classical Latin corpus at Packham Humanities Institute.

Notes: the Packham corpus does not distinguish between "I" and "j", or between "u" and "v". Word breaks incude spaces, periods, commas and question marks.

probability of the second character is over 20 percent.

At this point we could again raise the possibility that what appear to be two-glyph strings are not in fact composed of two separate glyphs. In chapter 3, we wondered whether the extraordinary combination ꝉo {4o} was in fact a single glyph. If it were, that would in turn suggest that the few instances of ꝉ {4} unaccompanied by o {o} were unfinished glyphs or misspellings. We could extend the same idea to other frequent combinations such as 8�unknown {89}, cc�unknown {C9} and 7�unknown {79}. It could be, for example, that 8�unknown {89} is not an 8 {8} followed by a �unknown {9} but has a meaning all of its own. Is it a coincidence that, as we observed in chapter 3, 8ꝝ was a medieval alchemical symbol which meant "mineral salt"? Alternatively, the terminal �unknown {9} could be an abbreviation, as in medieval Latin (of which, more later).

To explore these ideas would require a deep dive into the corpora of other medieval languages, and plenty of computing power, both of which for the moment are beyond the scope of this book.

OTHER "EVENTS"

The observation of a single glyph or space is not the only possible definition of a Markov event. We can also define an event as the observation of a glyph string, of any length we choose: two glyphs, three glyphs or more, not necessarily delimited by spaces, uncertain spaces, or line breaks. In this case, for, any given string, we can assess the

probabilities of its being followed by the same string, or another string, or a single character.

Table 7.3. Voynich manuscript, v101 transliteration: a Markov matrix of the ten most common two-character strings, and the frequencies of the following character

two-character string			third character									
			space	o	9	a	c	1	e	8	h	y
glyphs	v101	count	.	o	9	a	c	1	e	8	h	y
9 + space	9.	12,837	0%	20%	3%	1%	0%	13%	6%	8%	4%	2%
space + o	.o	7,210	1%	0%	0%	0%	1%	0%	20%	2%	28%	7%
oh	oh	5,863	1%	7%	6%	31%	26%	9%	0%	0%	0%	0%
space + 1	.1	5,811	0%	26%	3%	4%	43%	0%	1%	3%	2%	0%
oe	oe	5,406	57%	3%	4%	3%	0%	4%	0%	3%	9%	0%
4o	4o	5,250	1%	0%	0%	1%	1%	0%	5%	2%	57%	0%
e + space	e.	5,151	0%	20%	3%	3%	0%	23%	4%	10%	6%	1%
89	89	4,886	85%	0%	0%	0%	0%	0%	0%	0%	0%	0%
space + 4	.4	4,816	0%	97%	0%	0%	1%	0%	0%	0%	0%	0%
y + space	y.	4,677	0%	24%	4%	19%	0%	21%	1%	4%	1%	0%

Source: author's analysis.

probability of third character is over 20 percent

Again, we can give a simple illustration, using the ten most common two-character strings in the Voynich manuscript (including both glyphs and spaces), and examining what the following character is likely to be. The probabilities are summarized in Table 7.3.

This table shows that the patterns of the Voynich manuscript, which we saw at the two-character level, continue at the three-character level. If

anything, at this level they are even more pronounced. To cite just a few examples:

- The strings {oe} and {89} are very likely to be followed by a space; these strings are likely to be the ending of a Voynich "word."
- A terminal {9} or {e} is often followed by a string that begins with {o}; these are examples of the phenomenon which baffled Capt. Currier: that the ending of one word seems to influence the beginning of the next "word."
- An initial {1} is often followed by {o}.
- The string {oh} is often followed by {a} or {c}.
- The string {4o}, which is extraordinary in its own right, is very often followed by {h}.

and so on …

We begin to understand that the Voynich manuscript has deep structure. There is predictability at the level of three characters, and more. The interested reader is invited to refer to our Appendixes 2.1 through 3.4, where we look at strings of two and three glyphs, to understand how deep this structure goes.

Whatever the level at which we choose to search for the building blocks of the Voynich manuscript, we can construct a Markov matrix at that level. Given sufficient data, time, and computing power, we can compare the structure of the Voynich manuscript with that of any natural language. At the levels of one, two and three glyphs (as illustrated in the appendixes), this author has only attempted the comparison with six selected medieval languages. For the interested reader, comparisons of longer strings, and with other languages, could be productive areas of research.

In summary, by means of Markov models we can quantify and characterize the patterns within the Voynich manuscript: and we know from Currier's work and from our own analysis that these patterns are deep and pervasive. Through this kind of analysis, we might hope to find to what extent, if any, the Voynich manuscript resembles any natural language; and

if we find any resemblance, we could begin to attempt a transliteration. At least, we might hope for Markov models to tell us which natural languages are worth exploring, and which are not.

In modern times, in particular since the 1976 Voynich seminar, many scholars have examined the glyphs and the glyph strings from a mathematical perspective. Some have explicitly used Markov models; others have used code-breaking programs; others have used exotic mathematics such as fractal dimensions. In the rest of this chapter, we shall set out a summary of the modern search for patterns in the Voynich manuscript.

D'IMPERIO

Mary D'Imperio, in her book *An Elegant Enigma* and in her three technical articles on the Voynich manuscript, did not once mention Markov or the mathematical concepts that bear his name. However, it is clear that she was preoccupied with the structure and patterns of the manuscript, in the sense to which we have alluded above. In one of her articles, she wrote,

> … the text has a very clear and consistent structure … The occurrence of words within lines and symbols within words exhibits the operation of orderly rules … Certain sequences of symbols recur in similar parts of words consistently; some symbols regularly occupy preferred positions at the beginnings, middles, and ends of words, and at the beginnings and ends of lines; some symbols appear frequently before or after other symbols. and rarely elsewhere.[4]

Her comments were focused on the sequential nature of the text: the pervasive phenomenon that glyphs could predict the following glyphs, even to the extent of jumping across a "word" break. It was similar to, though vastly more complex than, the phenomenon that impelled Andrei Markov in his study of *Eugene Onegin*.

D'Imperio's approach to the Voynich structure was based on a program called PTAH, to which we referred in Chapter 6. Her descriptions of the program sound very much like a Markov model. Her whimsical analogy for the program had exactly the characteristics of a Markov chain:

- A hypothetical animal has five possible "states": eating, sleeping, hunting for food, sitting still, and scratching for fleas.
- Each "state" has some probability of being followed by one of the five "states," including the same state or any of the other four.
- Although we cannot know the animal's intentions, we can observe the animal for as long as we want, and see what it does, and in what sequence.
- With a sufficient number of observations, we can construct a mathematical model which, for any given state of the animal, predicts its subsequent states.

Applying this analogy to the Voynich glyphs, D'Imperio defined five possible "states" which a glyph could have, defined by reference to its position in the "word." She took a large chunk of the Voynich manuscript and fed it to the PTAH program. PTAH crunched the text iteratively until it had classified each of the Voynich glyphs into one of the five states. D'Imperio used Prescott Currier's transliteration alphabet, which had only twenty-six glyphs plus the word break and the line break. In order to place D'Imperio's work in the context of this book, as best we could, we mapped this alphabet to the v101 keyboard assignments.

D'Imperio's paper seemed to identify seven, not five, states for the Voynich glyphs, and did not provide explicit definitions of the states, although we can guess that they were probably as follows:

separators	a space or a line break; possibly an isolated glyph
beginners-1, -2	predominantly occur as the first glyph of a "word"

post-beginners-1	predominantly occur between the first glyph and the middle of a “word”
middles	predominantly occur at or near the middle of a “word”
pre-enders	predominantly occur between the middle and the last glyph of a “word”
enders	predominantly occur as the last glyph of a “word”

Table 7.4: Mary D’Imperio’s five states for Voynich glyphs, and corresponding glyph counts in the v101 transliteration, based on position in “word”

		associated Voynich symbols				author’s count of glyphs in v1			
state	label	D’Imperio glyph	Currier alphabet	v101 glyph	v101	total	initial	internal	final
1	beginners-1		O	o	o	25,043	8,598	14,883	1,366
			A	a	a	14,496	2,714	11,717	55
			V	f	f	410	118	261	25
				u	u				
4	beginners-2, post-beginners-1		E	e	e	10,652	1,216	2,994	6,172
			F	h	h	9,919	1,352	8,419	123
			R	y	y	6,717	261	978	5,200
			P	k	k	6,118	1,022	4,977	96
			4	4	4	5,408	5,372	30	0
			2	s	s	2,819	916	364	1,097
				t	t				
				T	T				
			N	n	n	1,770	0	29	1,739
			J	p	p	989	1	29	940
			B	g	g	879	465	391	17
				j	j	564	63	476	24
			D	N	N	132	0	52	80

5	middles		C	c	c	13,750	126	13,519	97
				cc	C	2,844	34	2,785	24
			S	cc	1	11,052	6,338	4,665	19
			Z	c̊c	2	4,065	2,970	1,056	10
				c̊c	3				
				c̊c	%				
				c̊c	!				
			Y	c̊c	F	35	18	17	0
3	pre-enders		8	8	8	13,208	3,748	8,677	666
				8	7				
				8	6				
				8	&				
			X	c̊c̊c?	L?	42	9	33	0
			Q	c̊c?	K?	910	489	411	5
2	enders		9	9	9	17,436	1,619	436	15,066
			M	ıı0	m	4,112	0	39	4,071
			T	ıı2	z	511	0	80	431
			U	ıı2	Z	154	1	16	136
			L	ıı8	q	11	0	7	4
			0	ı ıı2?	iZ?	0	0	0	0

Source: D'Imperio, Mary E., (1978?), p.72; correspondences with v101 alphabet and glyph counts by author "% matching label" is highlighted if percentage exceeds 40 percent.

On the basis of these assumptions, we tested D'Imperio's classifications of the glyphs against the actual counts in the v101 transliteration, as shown in Table 7.4. Having only the find-and-replace function of Microsoft Word at our disposal, we could identify internal glyphs, but we were unable to make the fine distinctions that the PTAH program made between beginners-1, post-beginners-1, beginners-2, middles and preenders. Our observations, with reference to the more common glyphs, were as follows:

- The PTAH program correctly identified {9}, {m}, {z} and {Z} as ending glyphs, while missing {p}, {n}, {N} and possibly {y}.
- The PTAH program correctly identified {c}, {C}, {k}, {8} and its variants, {g} and {j} as internal glyphs, while missing {o} and {a}.

The PTAH program was not very successful at identifying beginning glyphs; it missed {1}, {2} and all its variants, {4} and {K}, as well as identifying {o} and {a} as beginners, whereas these are predominantly internal to glyph strings.

Nevertheless, D'Imperio provided substantial statistical confirmation, with what looked very much like a Markov model, of the conclusions that William Friedman and John Tiltman had reached from their earlier, largely manual and intuitive, studies of the Voynich manuscript:

> The Voynich symbols do indeed fall into well-defined classes associated with beginnings, middles, and ends of words. In addition, there is a mechanical, regimented quality about the picture we see here - an appearance of surprising orderliness, a highly limited and regular behavior and a resultant degree of predictability.[5]

That was in 1978. There, as far as we can tell, the matter rested for over twenty years. Only in the 2000s would scholars, armed with twenty-first-century computing power, return to the structure of the Voynich manuscript. We shall report on their most important advances, as we perceive them, in the following sections of this chapter, and in Chapter 14.

But before we take our leave of Mary D'Imperio, we can report on our own attempts to advance her ideas. With no more than the power of Microsoft Word, we can have another shot at a classification of the glyphs, and possibly draw some inferences as to which of them represent natural letters, and which do not.

Our approach was as follows:

- import the v101 transliteration into Microsoft Word (for the moment, ignoring the differences between Currier's Languages A and B)
- convert all ASCII strings for "rare glyphs" to the Unicode symbol ֍
- convert all instances of {.} and {,} to spaces
- convert all line breaks to spaces (so that a glyph which started a line would be identified as the initial glyph of a "word," and a glyph which ended a line would be identified as the final glyph of a "word")
- convert all instances of the glyph string †o {4o} to the Unicode symbol ④ (since by now we were convinced that this string was in fact a single glyph); this had the effect of reducing the total glyph count to 153,682
- retain † {4} and o {o} for all other occurrences of these two v101 glyphs.

This gave us a flat text file in which initial, final and internal glyphs could be identified, as well as isolated glyphs which looked like single glyph "words." Page f002r, for example, now looked like this:

h98an9 9g1oe 8am ok1ae 9j1am Hoes9 8oy1oy9 1hay s 2oy K9 ֍ ④kam Kc9 9 1oy 19 989 1am ֍aI89 1ko8 89 G9 8aes 1 okam 8 oko1oy ae 2o8am 1oe 8aN 9k1am 8aN sam 8an8 8hoe soy 9koe89 81oe 819 K9 2oy H9 8am9 1oe 8aN h98an 2am ④9 s 2oe fo8aN 9h2 oe 2C9 8aIe89 8es 2o hoe 2C9 ④hc9 9ho89 so 1oe 9h9 8an 8aZoe ֍h9 1oeam 2oe 2ch9 8am Kc9 hcoe s am ֍֍ 91an 8ae 19 8aeoy 2aN 8aN oes am 2C9 Hoy ohoe 19 1oy Koy 9oy aN 1aN s am 1ck9 19h9 sae 2o 9hC9 1c9 8am 1K9

We confined the analysis to glyphs which occurred at least 140 times in the manuscript; there were thirty-four such glyphs, and collectively, they accounted for 98.4 percent of the total glyph count. The counts and distributions of these glyphs are summarized in Table 7.5.

This table seemed to tell us the following:

- Our newly defined ⟙o ④ and the redefined ⟙ {4} practically always begin a "word."
- Nine glyphs— o {9}, ⟍⟍J {m}, ⟍J {n}, ⟍9 {p}, ⟍Y {z}, q {(}, ⟍⟍J {M}, 9 {*} and ⟍⟍Y {Z} —are usually at the end of a "word."
- On the infrequent occasions when one of these eleven glyphs appears to be internal to a "word", we could well believe that the "word" is actually two or more "words," and that the glyph in question is actually an initial or final glyph of one of these "words."
- In natural European languages, there are no characters that invariably begin or end a word, unless they are punctuation marks, or certain abbreviations as in medieval Italian or Latin.
- Therefore, we should regard these eleven glyphs as "markers" of some kind, either punctuation or abbreviations, and we should exclude them from the glyph count and from all of our attempts at transliteration.
- In particular, we should calculate glyph frequencies only after the "marker" glyphs have been excluded (as shown in the right-hand column of Table 7.5).

Table 7.5. Voynich manuscript, v101 transliteration: counts of the most common glyphs, by position in "word"

glyph	v101	glyph count					percentage of count			
		all	initial	internal	final	isolated	initial	internal	final	is
total		153,682	*38,669*	*74,305*	*38,669*	*2,039*	25.2%	48.3%	25.2%	
total excl. "markers"		115,845	*31,526*	*72,656*	*10,437*	*1,226*	27.2%	62.7%	9.0%	

০, other	o, other	19,778	8,598	9,618	1,366	196	43.5%	48.6%	6.9%
৯	9	17,436	1,619	436	15,066	315	9.3%	2.5%	86.4%
ɑ	a	14,496	2,714	11,717	55	10	18.7%	80.8%	0.4%
c	c	13,750	126	13,519	97	8	0.9%	98.3%	0.7%
৫	1	11,052	6,338	4,665	19	30	57.3%	42.2%	0.2%
৪	e	10,652	1,216	2,994	6,172	270	11.4%	28.1%	57.9%
৪	8	10,343	3,247	6,413	584	99	31.4%	62.0%	5.6%
h	h	9,919	1,352	8,419	123	25	13.6%	84.9%	1.2%
২	y	6,717	261	978	5,200	278	3.9%	14.6%	77.4%
k	k	6,118	1,022	4,977	96	23	16.7%	81.4%	1.6%
৭০	4o	5,265	5,072	24	3	166	96.3%	0.5%	0.1%
ৱ	m	4,112	0	39	4,071	2	0.0%	0.9%	99.0%
৫	2	3,359	2,449	878	8	24	72.9%	26.1%	0.2%
८८	C	2,844	34	2,785	24	1	1.2%	97.9%	0.8%
৪	7	2,714	475	2,150	73	16	17.5%	79.2%	2.7%
২	s	2,702	883	310	1,077	432	32.7%	11.5%	39.9%
ৱ	n	1,770	0	29	1,739	2	0.0%	1.6%	98.2%
৪	p	989	1	29	940	19	0.1%	2.9%	95.0%
৫	K	910	489	411	5	5	53.7%	45.2%	0.5%
g	g	879	465	391	17	6	52.9%	44.5%	1.9%
৫	H	866	201	657	6	2	23.2%	75.9%	0.7%
ɑ	A	768	267	420	67	14	34.8%	54.7%	8.7%
j	j	564	63	476	24	1	11.2%	84.4%	4.3%
৫	3	516	384	129	1	2	74.4%	25.0%	0.2%
ৱ	z	511	0	80	431	0	0.0%	15.7%	84.3%
৫	5	409	277	129	2	1	67.7%	31.5%	0.5%
৭	(	402	55	7	331	9	13.7%	1.7%	82.3%
ι	i	264	0	257	7	0	0.0%	97.3%	2.7%
f	f	233	96	125	8	4	41.2%	53.6%	3.4%
u	u	177	22	136	17	2	12.4%	76.8%	9.6%

⁞⁞	M	174	0	1	170	3	0.0%	0.6%	97.7%
⁞	*	164	0	7	145	12	0.0%	4.3%	88.4%
⁞	Z	154	1	16	136	1	0.6%	10.4%	88.3%
⁞, other	4, other	143	134	3	0	6	93.7%	2.1%	0.0%

Source: author's analysis.

"marker" glyph, initial

"marker" glyph, final

These ideas will be elaborated in Chapter 8, wherein we will test the modified glyph frequencies against the letter frequencies in selected European languages.

REDDY AND KNIGHT

To this author's mind, the most concise and at the same time comprehensive examination of the Voynich "word" structure is the 2011 paper by Sravana Reddy of the University of Chicago, and Kevin Knight of the University of Southern California: "What We Know about the Voynich Manuscript."[6]

Reddy's and Knight's paper was addressed to an audience of scholars in linguistics, text analysis and machine learning. They had worked with Prescott Currier's transliteration, which had only thirty-seven distinct glyphs as opposed to Claston's seventy-one, and therefore they did not have to worry about the minute variations that Claston felt obliged to distinguish. The content of their paper was couched in technical terms; the unwary reader would need to know, for example, that HMM stands for Hidden Markov Model and that EM stands for Expectations-Maximization.

This author, who will admit to being a mathematician, has extracted the following takeaways from this paper with regard to Reddy's and Knight's views of the glyphs and the glyph strings:

- Four glyphs were likely to represent vowels: o {o}, a {a}, c {c}, and 9 {9}.
- The script might be an *abjad* (like Arabic or Hebrew script) in which short vowels are omitted and only consonants and long vowels are retained. In particular, the distribution of word lengths resembled that of their sample from a Buckwalter transliteration[7] of the Holy Quran.
- The "gallows" glyphs "are not uppercase versions of other letters, and … they are contextually similar to one another."[8] [This appeared to mean that a gallows glyph could be replaced by another gallows glyph and the result would generally be another Voynich "word."]
- There was no sign of punctuation as it was normally understood. By this, they meant that they had found no final glyph which could consistently be removed from a "word" and would still leave a "word." [This author thinks otherwise.]
- The sequencing of the glyphs was more predictable than that of letters in modern natural languages, in the sense that many glyphs were likely to be followed by specific other glyphs. The predictability of the Voynich glyphs was very similar to that of Chinese Pinyin (a phonetic transliteration of written Chinese characters).
- There were signs that the "words" had a structure consisting of prefix + stem + suffix.

Reddy and Knight expressed the following conclusions with regard to the manuscript as a whole:

- [The] "pages probably do have topics but are not independent of one another."[9] [This seemed like a restatement of the consensus that the manuscript is made up of themed sections; but it also seemed to state that individual words tended to cluster on specific pages. In other words, pages had characteristic vocabularies.]

- They confirmed Prescott Currier's observations that "there are patterns to lines on the page that are uncharacteristic of prose."[10]
- The pages were mostly in a coherent order, but "larger chunks of pages were probably reordered."

They concluded that the Voynich manuscript had some features which were like those of natural languages, and others that were not; but that the manuscript was not necessarily a hoax (a hypothesis that we examined in some detail in Chapter 4).

This author would add one observation with regard to *abjad* languages. In such a language, since the short vowels are omitted from the written language, any short string of letters is quite likely to make a word. For example, here are five strings of three numerals, all created by the random number function in Excel (in other words, without any human intervention):

15 20 18 16 18 15 17 6 26 17 3 23 27 20 3

We now map these numbers to the standard modern Arabic alphabet, which is ordered as follows:

1	2	3	4	5	6	7	8	9	10
ا	ب	ت	ث	ج	ح	خ	د	ذ	ر
11	12	13	14	15	16	17	18	19	20
ز	س	ش	ص	ض	ط	ظ	ع	غ	ف
21	22	23	24	25	26	27	28		
ق	ك	ل	م	ن	ه	و	ي		

This produces the following Arabic letter strings (reading from right to left):

ضفع	ظتل	ظحه	وفت	ضفع
fulfilled	shadow	smash it	bite	frog

Thus, from random numbers (which by definition, have no meaning), we were able to produce five perfectly good words or phrases in Arabic.

MONTEMURRO AND ZANETTE

In 2013, there appeared what we think was the first published scientific paper focusing on the "words" of the Voynich manuscript rather than single glyphs; and also, the first to test mathematically whether each thematic section contained "words" that represented keywords for that section.

This paper was titled "Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis"[11] by Marcelo A. Montemurro of the University of Manchester, England, and Damián H. Zanette of the Consejo Nacional de Investigaciones Cientificas y Técnicas in San Carlos de Bariloche, Argentina.

The two researchers used the following source documents:

- Jorge Stolfi's transliteration of the Voynich manuscript in the EVA alphabet,[12] reordered in order to make each thematic section a set of consecutive pages.
- *Confessiones* by Augustine, Bishop of Hippo (classical Latin, 141, 421 words)[13]
- *On the Origin of Species* by Charles Darwin (modern English, 200,470 words)
- *The Records of the Grand Historian* by Ssu-Ma Ch'ien (Chinese, about 526,500 Chinese characters)[14]
- a document in the Fortran programming language, consisting mainly of the test driver code for the BLAS (Basic Linear Algebra Subprograms)
- a document in genetic symbols, namely the genetic sequence of the yeast *Saccharomyces cerevisiae*[15] (12,156,677 base pairs)

Montemurro and Zanette were looking for the frequencies of Voynich "words" in their context. They explained:

> … the specific way in which words are distributed can be used to distinguish statistically different parts of a text. For instance, a word that only appears in one specific chapter of a book is a perfect tag for that chapter, i.e., if that word is encountered, one immediately knows with certainty which chapter is being scanned.[16]

Conversely, a "word" that was uniformly scattered throughout the manuscript would convey little or no information about the section where it was found.

In any document (whether the Voynich manuscript or another) the authors expected that the information content of words would depend on the size of the section under study. If the section were very small (say, ten words), they would not expect much information in a word. If the section were very large (for example 10,000 words), the words would be more uniformly distributed and would not convey much information. But somewhere in between, there would be a point at which the information content reached its maximum. They called this point the "optimum scale."

One of their key tasks was to see whether the "optimum scale" of the Voynich manuscript was similar to that of the other five documents.

Without getting into the mathematics of how one calculates the information content of a word, we reproduce their result in figure 7.2.

In summary, this graph says that in the Voynich manuscript, the information content of "words" reached its maximum when the "words" were observed in chunks of about 800 "words."

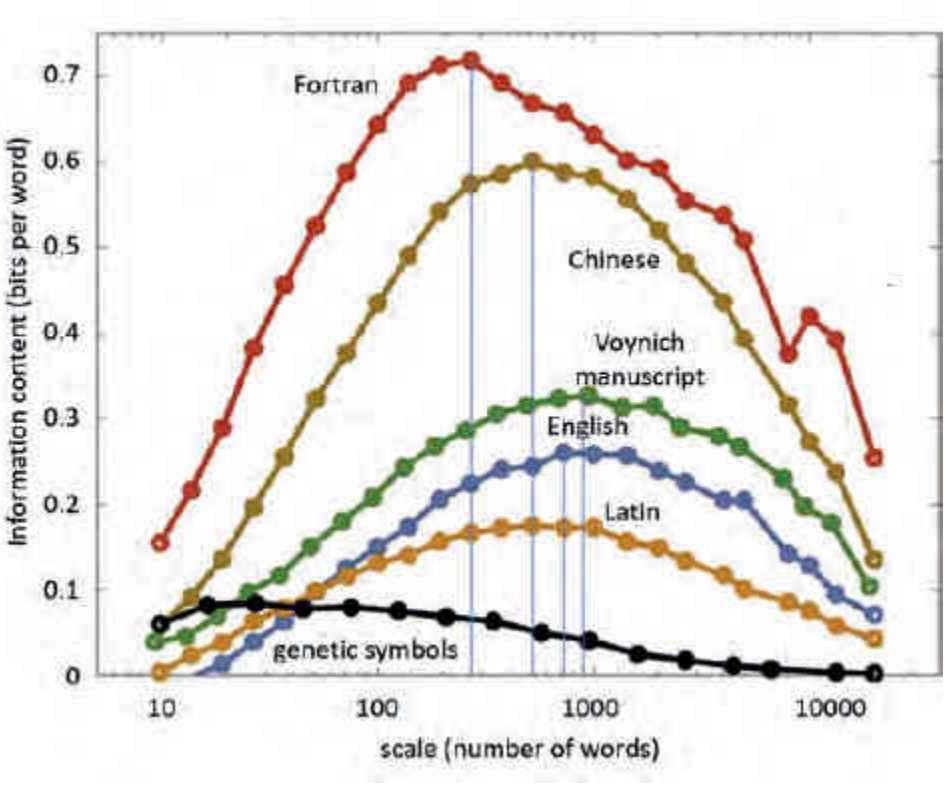


Figure 7.2. The Voynich manuscript and five other documents: information content of words, measured against scale of observation. *Montemurro and Zanette (2013), additional graphics by author*

We note that the "optimal scale," at 800 words or so, was a much-smaller chunk of the Voynich manuscript than the "thematic sections," which have conventionally been viewed as the "chapters" of the manuscript. The "herbal section," for example, has 129 pages and 11,687 "words."[17]

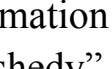
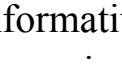
However, their important finding was that the manuscript was more similar to the English-language document, where the "optimal" scale was about 720 words, than to the documents in other natural languages (Latin and Chinese), and very dissimilar to the documents in Fortran and genetic symbols. Montemurro and Zanette commented that:

> … the overall statistics of word distribution over the text in the Voynich manuscript [are] comparable with that of real human languages.[18]

Turning to the links between the Voynich "words" and the thematic sections where they occurred, the authors found that the frequencies of "words" were observably different from one section to another. Certain sections were similar to each other: not just on the basis of their illustrations (which was how researchers had conceived the sections in the first place), but on the basis of their word frequencies. They concluded:

> The strongest link occurs between the Pharmacological and Herbal sections…. our analysis establishes, for the first time, a link between these two sections on the sole basis of the linguistic structure of the text.[19]

This was the first scientific evidence that the illustrations and the accompanying text were related: that the text was not merely an arbitrary addition to a book of illustrations. If that were so, we might hope to find among the "herbal" pages, for example, a glyph string that represents the name of a plant. Then we could derive some justification for looking for a "crib" among the herbal pages: a task on which we shall report in Chapter 9.

Finally, Montemurro and Zanette reported on the "words" with the highest information content. At the "optimal scale," that "word" was  {2c89}, or "shedy" in the EVA alphabet; at the scale of the thematic section, the most informative word was  {8am}, or "daiin" in EVA. They seemed to be saying that these "words" were, in some sense, keywords; but keys to what, their paper did not say.

However, we can try to reproduce their analysis. We have the text of each of the seven sections, imported into Microsoft Word, and we can count the most-common "words" with a variety of software, such as WordCounter. For any given section, the "words" which are unique to that section's "top twenty" can possibly be interpreted as keywords for that section. Without further ado, the results are summarized in Table 7.6.

Table 7.6. Voynich manuscript, v101 transliteration: most frequent "words", and key "words" in the major thematic sections

Voynich manuscript	herbal	text

glyphs	v101	count	% of 1st "word"	glyphs	v101	count	% of 1st "word"	glyphs	v101	co
ꬹꬰꬱ	8am	739	100.0%	ꬹꬰꬱ	8am	393	100.0%	ꬰꬱ	am	3
ꬰꬱ	am	739	100.0%	ꬻꬷꬶ	1oe	222	56.5%	ꬰꬲ	ay	2
ꬷꬶ	oe	710	96.1%	ꬲ	s	209	53.2%	ꬰꬶ	ae	2
ꬰꬲ	ay	489	66.2%	ꬰꬱ	am	161	41.0%	ꬷꬶ	oe	2
ꬷꬲ	oy	440	59.5%	ꬻꬷꬲ	1oy	159	40.5%	ꬶ	e	1
ꬲ	s	432	58.5%	ꬷꬲ	oy	135	34.4%	ꬻꬻꬹ	1c89	1
ꬻꬷꬶ	1oe	380	51.4%	ꬷꬶ	oe	123	31.3%	ꬹ	9	1
ꬻꬻꬹ	1c9	375	50.7%	ꬹ	89	121	30.8%	ꬻꬻꬹ	1c9	1
ꬻꬻꬹ	1c89	360	48.7%	ꬻꬹ	19	121	30.8%	ꬹꬰꬱ	8am	1
ꬰꬶ	ae	361	48.8%	ꬹ	9	109	27.7%	ꬷꬲ	oy	1

balneology / biology | pharmaceutical | cosmolo

glyphs	v101	count	% of 1st "word"	glyphs	v101	count	% of 1st "word"	glyphs	v101	co
ꬷꬶ	oe	297	100.0%	ꬹꬰꬱ	8am	72	100.0%	ꬰꬲ	ay	5
ꬻꬻꬹ	1c89	151	50.8%	ꬷꬶ	oe	68	94.4%	ꬰꬱ	am	4
ꬾꬵꬰꬱ	4ohan	150	50.5%	ꬲ	s	67	93.1%	ꬷꬶ	oe	4
ꬾꬷꬶ	4oe	123	41.4%	ꬰꬱ	am	55	76.4%	ꬷꬲ	oy	4
ꬼꬻꬹ	2c89	109	36.7%	ꬷꬲ	oy	50	69.4%	ꬰꬶ	ae	3
ꬾꬵꬻꬹ	4ohc89	105	35.4%	ꬻꬷꬶ	1oe	47	65.3%	ꬹ	9	2
ꬾꬵꬰꬶ	4ohae	97	32.7%	ꬻꬻꬷꬶ	1coe	42	58.3%	ꬷ	o	2
ꬻꬻꬹ	1c9	95	32.0%	ꬹ	9	31	43.1%	ꬹꬰꬱ	8am	2
ꬾꬵꬻꬻꬹ	4ohC89	80	26.9%	ꬷꬵꬷꬶ	ohcoe	28	38.9%	ꬹꬰꬲ	8ay	2
ꬾꬵꬰꬱ	4oham	76	25.6%	ꬹꬰꬱ	7am	27	37.5%	ꬹꬰꬶ	8ae	2

Source: author's analysis; case-sensitive word counts from Yasumasa Someya, *Do-It-yourself Word Counter*.

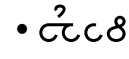
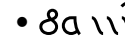
most frequent ten "words" in manuscript

key "words" for biology section

key “words” for pharmaceutical section

It is clear that all seven sections share some glyph strings, usually very short strings. In addition, each of the sections has some unique glyph strings which suggest “keywords” for that section, although at this point, we have no idea what these glyph strings mean, if they mean anything. These results seem to say that the manuscript is an integrated whole, with a shared vocabulary but with specific vocabularies for each of the thematic sections.

As for the words that Montemurro and Zanette identified as keywords:

-  {2c89} was the eighteenth most frequent “word” in the manuscript, but the fifth most frequent in the “biology / balneology” section; it seemed thereby to represent a keyword for that section;
-  {8am} was the (equal) most frequent “word” in the whole manuscript, and also the most frequent “word” in the “herbal” and “pharmaceutical” sections; it seemed therefore to confirm a linguistic link between these two sections, independent of the visual similarities of the illustrations.

HAUER AND KONDRAK

In 2016, Bradley Hauer and Grzegorz Kondrak of the University of Alberta published a paper[20] on a new mathematical approach to the concept of encipherment of the Voynich manuscript.

This paper has particular importance because, to this author’s knwledge, it is the only paper from an academic institution which proposes a specific, if partial, glyph-to-letter mapping. The proposed mapping is to Hebrew.

If this mapping is valid, it could in principle yield a translation of the whole of the manuscript into Hebrew and thence into any other language. For this reason alone, the paper deserves meticulous scrutiny

Hauer and Kondrak started from the assumption that the text of the Voynich manuscript had been enciphered in one of two ways:

- With a monoalphabetic substitution cipher. This is the simplest of all ciphers, in which each character of the plaintext (possibly including spaces as well as letters) is mapped one on one to a character in the ciphertext. It is, as we recall from chapter 3, the cipher which William Legrand cracked in "The Gold Bug," and of which Sherlock Holmes made short work in "The Adventure of the Dancing Men." In such a cipher, the Italian words "aglio e olio" could become "xzrtfybycrtf," and the observant decipherer would notice that the string "rtf" was repeated but might not guess that "y" was a word break.
- With an anagrammed substitution cipher, in which each word in the plaintext retains its original length but the letters within the words are scrambled in some arbitrary way. In such a cipher, for example "aglio" could become "olaig" and the task of the decipherer would be to sift the 120 possibilities and return to the original "aglio."

Our understanding of Hauer and Kondrak's approach, as they described it, is as follows:

- They had examined three computational methods for identifying the source language of a document that had been enciphered with a monoalphabetic cipher, namely:
  - frequency analysis: comparing character frequencies in the ciphertext with letter frequencies in many natural languages (as did the fictional William Legrand and Sherlock Holmes, but only for English)
  - decomposition pattern frequency: an elaboration of frequency analysis, which looks at multiple occurrences of characters within words (for example "olio," however enciphered, will always have the same first and last letter)
  - trial decipherment: a further elaboration of frequency analysis, in which the ciphertext is mapped to the candidate language and then undergoes repeated swapping of letters to improve the similarity to readable text.

- They had devised a method for decoding a document that had been enciphered with an anagrammed cipher.
- They had tested their methods on the first line of the Voynich manuscript.

Their paper was framed in mathematical terms and drew upon Markov models. This author has no dispute with the mathematics. The interested reader can look up Hidden Markov Models, and Expectation Maximization, or take the math on trust. Let us then cut to the chase. On the Voynich manuscript, Hauer's and Kondrak's results were as follows:

- Using straight frequency analysis, Hauer and Kondrak found at least ten languages which had similar character frequencies to the Voynich manuscript. Their paper mentioned only four of the top ten: Mazatec (ranked #1), Mozarabic (#3), Italian (#8) and Ladino (#10).
- Using the decomposition pattern method, the top-ranked candidate languages were Hebrew (#1), Malay in Arabic script (#2), Standard Arabic (#3), and Amharic (#4).
- The trial decipherment method produced varying results depending how the programmers set the parameters, but in most of the trials, Hebrew and Esperanto were near the top.
- With regard to the anagram decipherment, Hauer and Kondrak concluded: "… if the VMS encodes a natural language text, the letters within the words may have been reordered during the encryption process." This seems to mean that the anagram approach did not work.

They concluded the paper with a translation of the first line of page f1r, a page which, according to both Currier and Davis, is in Language A. In this line, the Voynich glyphs (read left to right) were mapped to Hebrew letters (reading right to left). Their article reproduced the Hebrew text as follows (reading right to left):

המצות ועשה לה הכהן איש אלין לביחו ו עלי אנשיו

They conceded that this did not make a coherent phrase in Hebrew, but proposed the following free translation:

> She made recommendations to the priest, man of the house and me and people.[21]

We have four observations on Hauer and Kondrak's paper.

Firstly, as the authors made clear, and presumably in order to have the maximum number of candidate languages, they conducted the frequency analysis by reference to the *Universal Declaration of Human Rights* of 1948, which at the time had been translated into 380 languages. But this is a relatively short document: the English version has only 9,164 characters.[22] We could expect the letter frequencies of such a document to be different from that of the language as a whole. In the English version, which we have checked, this is the case: the letter frequencies are up to 1.5 percentage points different from that of a modern English corpus.[23] It would be instructive to repeat the frequency analysis with larger corpora, even if that meant reducing the number of candidate languages.

If this were to be done, it would be even better if the corpora were based on medieval documents: for example, the complete works of Shakespeare (884,647 words[24] and probably around 3.5 million letters). If medieval Italian were considered a candidate language, a good corpus would be Dante's *La Divina Commedia* (100,521 words and 444,962 characters);[25] an even better one would be the OVI corpus of medieval Italian (29,926,438 words).[26]

Secondly, as we observed earlier in this chapter, an attempted transliteration to an *abjad* language like Hebrew can produce what looks like meaningful text even if the underlying document has no meaning.

Thirdly, if we understood Hauer and Kondrak's mapping correctly, it includes the equivalences that we have set out in Table 7.7. However, the frequencies of the proposed Hebrew equivalents have little relationship to the frequencies of the glyphs in Language A.

For example, the v101 glyph {1}, which has a frequency of 8.6 percent, maps to the Hebrew letter צ, which has a frequency of 1.3 percent. The relatively rare v101 glyph {f} maps to the common Hebrew letter מ.

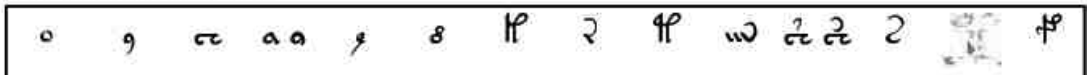
For the fourteen v101 glyphs which the authors proposed a mapping to Hebrew, and the corresponding Hebrew letters, the correlation coefficient is 37.4 percent. This is far lower than any of the frequency correlations that this author has calculated, between v101 glyphs and letters in modern or mediecal European languages.

In any case, the mapping can be tested by applying it to other text in Language A. As shown in Table 7.7, this author has done so, for the ten most frequent "words" in Language A only three of these yielded real (but short) Hebrew words. Interested readers who understand Hebrew have an opportunity for further research.[27]

Our fourth observation is more fundamental. If Hebrew was the underlying language of the Voynich manuscript, we have to imagine that the Voynich scribes were working from a written Hebrew document or documents. It would not be necessary for them to understand Hebrew; it would probably have helped if they did not. We can further imagine them mechanically transcribing the Hebrew letters into the script that they had been instructed to use. In this scenario, we envision a large board or tablet set up for their reference, with a column of letters in Hebrew and a corresponding column of glyphs in what we now know as the Voynich script. Again, it does not matter whether this was a script that was familiar to them and has now been lost; or whether it had been invented for the sole purpose of the project.

Table 7.7. Voynich manuscript, Currier transliteration (Language A): Hauer's and Kondrak's mapping of glyphs to Hebrew letters

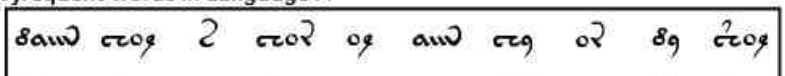
glyphs identified on page f001r, line 1

glyph														
Currier	O	9	S	A	E	D	F	R	P	M	Z	2	Q	V
v101	o	9	1	a, A	e	8	h	y	k	m	2, !	s	= K?	f
frequency	19.0%	9.8%	8.6%	8.2%	6.6%	6.2%	5.1%	4.7%	3.6%	3.1%	2.5%	2.3%	1.2%	0.1%

proposed mapping to Hebrew

letter	י	ו	צ	ה	ש	נ	ע	ל	כ	ד	א	ת	ב	מ
phonetic	yod	vav	tsadi	he	shin	nun	ayin	lamed	kaf	dalet	alef	tav	bet	mem
frequency	11.9%	10.9%	1.3%	8.8%	4.8%	3.4%	3.2%	6.7%	2.2%	2.7%	4.8%	5.8%	5.2%	5.3%

the ten most frequent words in Language A

"word"										
v101	8am	1oe	s	1oy	oe	am	19	oy	89	2oe

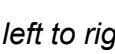
proposed mapping to Hebrew

"word"	דהנ	שיץ	ת	ליץ	שי	דה	וץ	לי	ונ	שיא
English	-	-	-	-	gift	-	-	to me	-	peak

Source: Currier keyboard assignments and Hebrew equivalents from Hauer and Kondrak (2016), p.84. Identification of corresponding glyphs by author. v101 equivalents from Glen Claston's v101 alphabet.

Note: the glyph which Claston classified as "rare" {} was identified by Currier as which he called Q, equivalent to {K} in the v101 alphabet.

Table 7.8. Voynich manuscript, page f1r, line 1, Currier transliteration: Hauer's and Kondrak's mapping to Hebrew

glyphs									
glyphs read from left to right									
v101 (L-R)	fa19s	9,hae	ay	Akam	2oe	!oy9	scs	9	hoy
Currier	VAS92	9FAE	AR	APAM	ZOE	ZOR9	QOR92	9	FOR
Hebrew	מהצות	ועהש	הל	הכהך	איש	אילו	בילות	ו	עיל
English[1]	from	and moth	cardamom	the dark	man	which	spending time	and	above
glyphs read from right to left									
v101 (R-L)	s91af	eah,9	ya	makA	eo2	9yo!	scs	9	yoh

Currier (R-L)	29SAV	EAF9	RA	MAPA	EOZ	9ROZ	29ROQ	9	ROF
Hebrew	תוצהמ	שהעו	לה	דהכה	שיא	וליא	תוליב	ו	ליע
English[1]	-	stayed	her	faded	record	-	-	and	-

Source: Currier keyboard assignments and Hebrew transliteration from Hauer and Kondrak (2016), p.84. v101 transliteration from Zandbergen, Rene, *The v101 transliteration file*. Note 1: from Google Translate. Image: *Beinecke Rare Book and Manuscript Library*.

However, a Hebrew document, then as now, would have been written from right to left; and it is universally agreed that the Voynich manuscript was written from left to right. Therefore, a scribe working from a Hebrew document would have had to write, on each line that he or she transcribed, the last Hebrew letter first.

In this case, we have to reverse every word on the Voynich line, and to reverse the order of the words on the line. As Table 7.8 shows, this changes completely the mapping to Hebrew, and therefore changes any meaning that the mapped words may possess.

None of this excludes Hebrew as the underlying language. Given the multifaceted and deeply analytical approach that Hauer and Kondrak brought to bear on the Voynich manuscript, there is surely scope for more work in a similar vein. At a minimum, their mappings of glyphs to letters, in Hebrew and in other languages, need to be distributed in full to researchers and tested against other pages of the manuscript.

ZELINKA, ZMESKAL, WINDSOR, AND CAI

In 2019, four authors—two from the Czech Republic and two from the USA—published an entirely new approach to the Voynich manuscript. The paper by Ivan Zelinka, Oldrich Zmeskal, Leah Windsor, and Zhiqiang Cai was titled “Unconventional Methods in Voynich Manuscript Analysis.”[29]

The authors had only one objective: to assess mathematically whether the Voynich manuscript was based on a natural language or was what they termed "historical fake text." We assume that their term "fake" meant simply any text generated randomly or by an algorithm, and therefore containing no meaning. Their approach was to develop an algorithm based on fractal geometry.

Some definitions are in order. Every object has a fractal dimension. This is a mathematical term, invented by Benoit Mandelbrot (he of the "Mandelbrot set," said to be the most complex object in the universe). Simply expressed, the fractal dimension of an object is a measure of its complexity.

The object most often used to explain fractal dimensions is a coastline. Seen from space, or on a map of the world, the coastline of any country looks quite smooth. One could imagine laying a string along it, and the length of the string would be the length of the coastline. On the other hand, a moment's thought (or a moment's zooming on Google Earth) will tell us that the closer we look at the coastline, the longer it is. From an airplane at an altitude of a few miles, we can see capes, promontories, fjords, bays, deltas, and river estuaries that were not visible on the map of the world. From a hill overlooking the sea, we can see rock pools, caves, and crevices that we could not see from the airplane. And if we are standing on the beach, we can see the coastline advancing and retreating with the waves. In short, the complexity of the coastline depends on the distance from which we view it.

Nature in general, and not just coastlines, is full of complexity. Benoit Mandelbrot wrote a book, *The Fractal Geometry of Nature*, to expound the concept that nature is complex at any scale. He used the term "fractal" to refer to an object which has this property: it exhibits complexity no matter from what distance we observe it. His eponymous "Mandelbrot set" is an object of infinite complexity, although it can be specified by mathematical equations that occupy one or two lines.

We will not attempt here to elaborate the mathematics that Zelinka and his colleagues used. For our purposes it is sufficient to report that they worked with four sets of data:

- the novel *The Old Man and The Sea* by Ernest Hemingway (in the original English, which has 134,822 characters including spaces,[30] and also in translations to 119 other languages)
- the English-language Bible, which in the King James version has 3,116,480 letters excluding spaces
- the Currier transliteration of the Voynich manuscript (we assume that they used the original Currier-D'Imperio version, which has 83,043 characters including spaces and line breaks, and which we assume is a partial transliteration of the manuscript)[31]
- a randomly generated set of 33,663 words.

Again, we cut to the chase. Zelinka and his colleagues came up with the following results.

In their first experiment, measuring the fractal dimensions of the Voynich manuscript, they found that its dimensions were within the boundaries of the 120 natural-language translations of *The Old Man and the Sea*. They concluded that the manuscript was based on a natural language.

In their second experiment, which they described as "the conversion of the Voynich manuscript into a complex network structure" (we will take their word for it), they compared the structure of the manuscript with that of random text. They explained this comparison as follows:

> Any logical text that carries some information, consists of letters, words and sentences that cannot repeat in an arbitrary order. Rather, due to the grammatical rules of the language, some repeated structures and text causalities must be observable. For example, some languages follow a subject-verb-object grammatical pattern, while others use subject-objectverb. If the text is converted into a network structure, then such a network must look different than a random-based [network].[32]

They found that, whether visualized graphically or computed quantitatively, the structure of the Voynich manuscript and that of random text were not similar. They concluded that:

> [The Voynich manuscript] is very likely non-fake text, [and is] written in [a] natural language.[33]

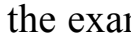
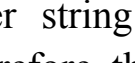
LAYFIELD, VAN DER PLAS, ROSNER, ABELA

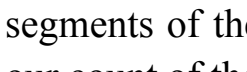
To bring us up to date, or nearly so, we must report on a new mathematical analysis of the Voynich manuscript that is based on "words," or glyph strings, rather than glyphs. In 2020, four researchers at the University of Malta—Colin Layfield, Lonneke van der Plas, Michael Rosner, and John Abela—published their paper "Word Probability Findings in the Voynich Manuscript."[34]

Layfield and his colleagues used the following data sets:

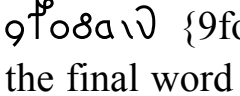
- an extract from the IVTFF (Intermediate Voynich Transliteration File Format) based on the EVA (Extensible Voynich Alphabet), from which "words" with uncertain glyphs or uncertain spaces were excluded; the residual text contained 36,249 "words"
- corpora of modern English, French, Hebrew, Korean, and Slovak, of unspecified size but apparently much larger than the Voynich manuscript.

For the purpose of comparisons between the Voynich manuscript and the five natural languages, they made the further decision to work only with strings of four to eight glyphs in Voynich, and words of four to eight letters in the natural languages. In the Voynich manuscript, these strings account for 84 percent of all the "words."

For any given word, they first established the "uniqueness point." They gave the example of the six-glyph string  {9fo8an} (which occurs only once in the whole manuscript). In their transliteration file, there was no other string which began with the first five glyphs  {9fo8a}. Therefore, the fifth glyph was the uniqueness point.

The next step was to segment each word up to the uniqueness point. If we understood their methodology correctly, the segments were simply single glyphs as defined in the EVA alphabet. If so, the cumulative segments of the string  {9fo8an} would be as follows (along with our count of their occurrences in the whole manuscript):

1.		{9}	(17,435 occurrences)
1-2		{9f}	(17 occurrences)
1-3		{9fo}	(4 occurrences)
1-4		{9fo8}	(2 occurrences)
1-5		{9fo8a}	(1 occurrence)

The heart of their analysis was a concept which they called “mean segment information.” An example may suffice. In the case of the “word”  {9fo8an}: segment 5 adds no new information: we know what the final word is going to be. For segment 4, the segment information is 33 percent (the additional glyph reduces the possibilities of the final word by 33 percent). For segment 3, the segment information is 19 percent, and so on. (We have simplified the explanation, since Layfield and his colleagues used the logarithms rather than the raw numbers shown above.)

The idea underlying this concept is that the more glyphs we see in a string, the more information we have about what the final string is going to be. To take an example from English: the string “canno” cannot predict the word “cannot,” because the string could also become “cannon.” We need the information in the final letter before we know the word. But the string “glyp” is practically certain to end up as “glyph.” If Voynich resembles a natural language, it should have some of the characteristics of a natural language in the extent to which the segments of a word predict the word.

Without getting deeply into the mathematics, we can summarize the findings of Layfield and his colleagues as follows:

> The [Voynich manuscript] follows the same pattern as the other five natural languages in that it shows that less probable words

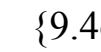
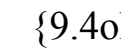
> contain more informative segments…. not only at word level but also at the level of segments, [the Voynich manuscript] shares characteristics with other natural languages…. we showed more support for the claim that the [Voynich manuscript] is written in a natural language and therefore is not a hoax.[35]

This author would observe only that these researchers, like all others who referred to Voynich "words," necessarily assumed that glyph strings delimited by spaces or line breaks were "words." It would have been a major effort to assume otherwise and to pursue the consequences. It is a natural and intuitive assumption; but it could be wrong.

In Captain Currier's presentation to the 1976 seminar, he emphasized the extraordinary phenomenon that the endings of some "words" seemed to predict the beginnings of the following "words." We quote from the transcript:

> Currier: "In no language [that] I know of does the suffix of a word have anything to do with the beginning of the next word."
>
> Question: … "How do you account for the full-word repeats?"
>
> Currier: "That's just the point—they're not words!"[36]

Our own analysis backs up what Currier was saying:

- The third most common 3-character string is {9.4}, with 3,405 occurrences.
- The single most common 4-character string is  {9.4o}, with 3,314 occurrences.
- The single most common 5-character string is  {9.4oh}, with 1,976 occurrences.
- And there are many other strings of three characters or more, with a space embedded, which occur thousands of times.

However, in the English language there are common strings of letters which straddle the spaces between words. In English the most frequent final letters are “e,” “t,” “s,” and “d”; the most-common three-letter words are “the” and “and.” According to the Corpus of Contemporary American English, which contains over one billion words, the most-common five-character strings containing these sequences are:

- “-t the-,” which occurs 6,503,524 times, in phrases such as “at the,” “that the,” and “about the”
- “-d the-,” which occurs 5,391,188 times, in phrases such as “and the,” “and then” and “and they”

The Voynich strings with spaces embedded may not be as bizarre as Currier thought; they may be no more than transcriptions of phrases equivalent (in some language) to “at the” or “and the.”

Nevertheless, the Voynich space may not be the beginning or end of a word. It may be a glyph, hiding in plain sight. We know from our computer keyboards and our telephones that in twenty-first-century technology, a space is a character. In Microsoft Word, if we press the ¶ key, we can see the space; otherwise, it is invisible, but it is still there.

The space occurs 32,475 times in the Voynich manuscript (34,906 times if we include “uncertain spaces”). Therefore, if it is a glyph, it is the most common one. In classical Latin (and we suspect, in most Indo-European languages), the space occurs more often than the most common letter. So, if we pursue frequency analysis, either we will come back to the view that the Voynich space is a word break; or if it represents a letter and not a word break, we will have to consider that it represents more than one of the most common letters in the underlying languages.

THE SPACE AS GLYPH

If we want to pursue this idea, we have to imagine that the Voynich scribes used the space consistently to represent something (other than a space) in the original documents.

As an entirely hypothetical example: if the Voynich manuscript had been based on the first printed edition of Dante's *La Divina Commedia*, the scribes would have read the first line as follows:

Nel mezo delcamin dinrã uita

which, in the modern representation, reads "Nel mezzo del cammin di nostra vita" ["In the middle of the pathway of our life"].

Let us imagine the scribes reporting to a project manager, who has defined the budget, schedule, and deliverables. Suppose that the project manager has given them instructions along the following lines:

> *Wherever you see a space in the original document, replace it with this glyph* (and the chances are that the glyph in question would be the ubiquitous o {o})
>
> *Wherever you see the following letter (or letters) in the original document, replace them with a space* (and the chances are, given the brevity of the Voynich words and the corresponding prevalence of spaces, that the project manager specified at least two letters: maybe "e" and "a," the two most common letters in medieval Italian).
>
> *For all other letters in the original document, use the mapping that I have written on this big board* (and the chances are that, given the frequencies of letters in medieval Italian, that the board would have looked something like the following:
>
> i = ⟨, o = a, n = c, r = cc, t = ⟨, l = ∂, s = ff, c = ?, d = ff, m = ⟨⟨ , u = f, z = f

Then the scribes would have written out, obediently and mechanically, the glyph strings shown in the second line of the image below:

This line appears to have six “words.” The average length of these “words” is 3.7 glyphs, which is close to the actual average for the Voynich manuscript. But, reading this line some centuries later, we would have no idea that the original spaces had mapped to a glyph and that, conversely, the Voynich spaces represented letters. Nor would we have any way to know that the original document had had five words with an average length of 4.8 letters.

Suppose that by a heroic feat of decipherment, we succeeded in mapping all the glyphs back to the Latin letters from which they came, except we did not guess that the o {o} was just a space. Suppose that we also assumed that the spaces that we saw between the glyphs had been spaces in the original document. Then our transliteration would read something like the following:

n lom zaod lc minodinr ouit.

It would need another heroic feat of imagination to see in this transliteration the first line of *La Divina Commedia*:

nel mezo delcamin dinrã uita.

For any reader who suspects that the space is a glyph, and wishes to pursue the implications, there is a rich field of research awaiting you.

WHERE NEXT?

From the research that we have attempted to summarize in this chapter, we leave with some degree of confidence that the Voynich manuscript possesses deep structure that, at various levels, resembles or at least is compatible with natural languages.

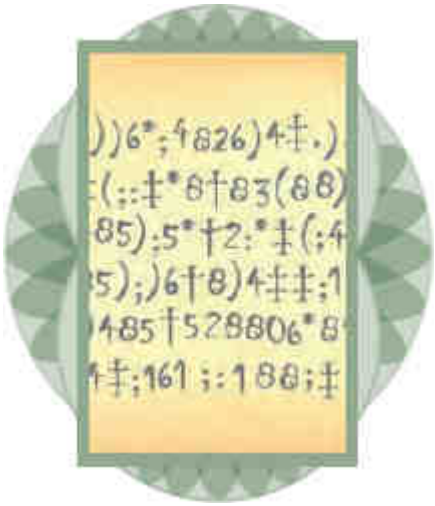
Although we have seen from chapter 4 that it is possible to create a Voynich-like text with an algorithm, we have found no evidence that the Voynich manuscript was so constructed. On the contrary, we have found

substantial scientific support for the idea that the manuscript was derived in some way from a document or documents in, as yet unidentified, natural languages.

What those languages might have been, will be the subject of the following chapters.

CHAPTER 8

THE GOLD BUG



Up to now, perhaps inspired by Dante's *La Divina Commedia*, we have been imagining the Voynich manuscript as a work that was possibly transcribed from documents in medieval Italian. Applying Occam's razor, it is the simplest hypothesis. The manuscript was written on parchment from the early fifteenth century, and we have learned that Wilfrid Voynich purchased it near Frascati in Italy. If the text did not originate in Italy, or was written later than the early fifteenth century, we need more complex assumptions about the provenance and travels of the manuscript. We may be permitted to wonder whether the nearly invisible signature on the first page is really that of Jacobus de Tepenecz; for if it is not, we have only a thirdhand report to support the proposition that the manuscript once resided in Prague.

If a similar document had been found in the archives of an English monastery, we might be forgiven for supposing that it had been written in English; or, to be more precise, transcribed from documents in English into the glyphs that we see today.

However, Prescott Currier gave us reason to believe that there was more than one underlying language. What we can do, and must do, now is to repeat the "Gold Bug" approach that we attempted in Chapter 5, but for Currier's Languages A and B separately. In other words, we must test whether Languages A and B[1] are statistically similar to European languages.

There is no harm in starting, as we did in Chapter 5, with medieval Italian as characterized in the OVI corpus. We can easily plot the frequencies of the most-common glyphs in Language A and Language B, as represented in the v101 alphabet, against the frequencies of the most-common letters in Italian. The results are illustrated in Figure 8.1.

It will perhaps be apparent from Figure 8.1 that Language B has a much better fit with medieval Italian than does Language A. Or if it is not visually apparent: in mathematical terms we can express the fit in terms of correlations: the frequencies in Language A and medieval Italian have a correlation coefficient[2] of 83.3 percent, whereas those in Language B and medieval Italian have a correlation coefficient of 97.7 percent.

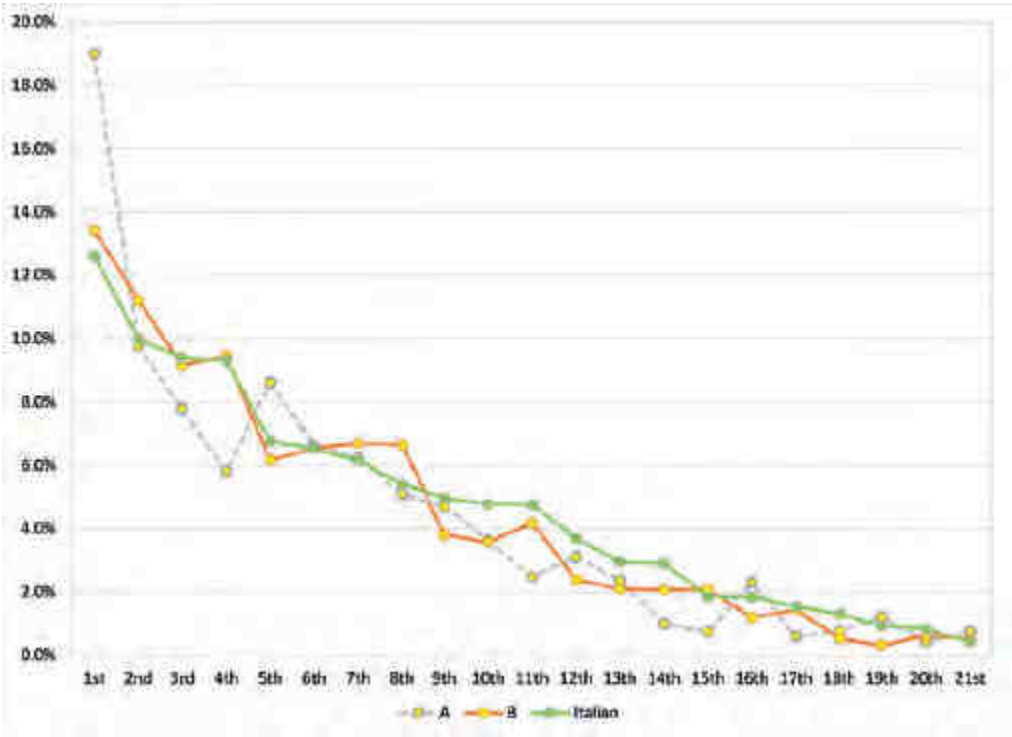


Figure 8.1. Voynich manuscript, v101 transliteration: frequencies of the twenty-one most common glyphs, in Currier languages A and B; and the frequencies of the twenty-one most common letters in medieval Italian. *Source: author's analysis*

For example, the dominance of the glyph o {o} is much less pronounced in Language B than in A; in this respect its frequency resembles that of the letter “e” in Italian.

This correlation raises the possibility (and it is no more than that) that, if there is any natural language underlying Language B, then medieval Italian could be that language.

CORRELATIONS

However, at this point it would be prudent to examine the correlations of Languages A and B, not only with Italian, but with a selection of European languages that might have been the basis of the Voynich manuscript. For this purpose, we drew upon the excellent website of Stefan Trost Media, which provides data on letter frequencies in twenty-nine modern European languages plus Hebrew, to which we added our own calculations of the letter frequencies in modern Hungarian, modern Romanian and medieval Italian, and Marsha Lynn Moreno’s estimates of the letter frequencies in medieval English.[3]

We noted that at least six of these languages, in modern usage, are heavily accented: Croatian, Czech, French, Galician, Portuguese, and Spanish. In the Voynich manuscript, there are several glyphs that look as if they might have accents: principally, they are cc {2}, the less common cc {3}, cc {5}, and the infrequent cc {!}, cc {#}, cc {%} and cc {+}, all of which look like accented variants of the glyph c {1}. It seemed possible that, if the Voynich manuscript was a transliteration from accented languages, the original accents might have been ignored or truncated. In these six European languages, therefore, we recalculated the letter frequencies without distinguishing accents: in French for example, combining the variants “e,” “é,” “è,” “ê,” and “ë.”

That gave a total of thirty-nine candidate languages: thirty-two European languages with accents, six European languages with accents ignored, plus Hebrew.

Our approach, for each of the Currier languages A and B, was to calculate the correlation of the glyph frequencies with the letter frequencies

in each of the target languages. In Microsoft Excel, this calculation is simplicity itself; it just requires application of the RSQ function. Our hypothesis was that the higher the correlation, the greater the probability that the candidate language was the one of the languages underlying the Voynich manuscript.

The results are summarized in Table 8.1.

These correlations tell us the following stories:

- They support Prescott Currier's thesis that Languages A and B are statistically different; and if they are not different languages, at least they have substantially different glyph frequencies.
- We have a wealth of choice. For Language A, there are at least eighteen languages (not counting variants without accents) in which the letter frequencies have a correlation of over 90 percent with the glyph frequencies; for Language B, there are at least twenty-six such languages.
- We can see that for Language B, as we had surmised earlier, medieval Italian is confirmed as a strong candidate for the underlying language, with a 97.3 percent correlation. We also see that Macedonian and at least ten other languages are equally good candidates; the differences in the correlations are statistically negligible.
- However, for Language A the target languages have an entirely different ranking. The top candidates, in statistical terms, are German, Irish, Norwegian, French, and Dutch.

We can see also that removing the accents from the target languages generally does not improve the correlation. In other words, if Languages A and B were originally mapped from accented languages, it looks as though the transcribers retained the distinctions between accents. If so, that would help account for the proliferation of glyphs in the v101 keyboard assignments.

However, our task has only just begun. We now need to test the hypothesis that the languages that are ranked highest in the correlations are indeed among the underlying languages of the Voynich manuscript. If they

are, then it should be possible to transliterate some passages from the manuscript into the target languages, glyph by glyph, and some of the transliterations should make sense.

Table 8.1. Correlations between glyph frequencies in Currier Languages A and B of the Voynich manuscript (v1010 and letter frequencies in selected European languages (plus Hebrew)

language	correlation with Language A
German	98.5%
Irish	98.4%
German (without accents)	98.2%
Norwegian	97.9%
French (without accents)	97.9%
Dutch	97.8%
English, medieval	97.4%
Irish (without accents)	97.3%
Danish	97.3%
French	96.1%
Spanish	93.5%
Spanish (without accents)	92.5%
Macedonian	92.5%
English	92.3%
Romanian (without accents)	92.2%
Russian	92.0%
Belarusian	91.9%
Portuguese	91.7%
Turkish	91.6%
Latvian	91.3%
Estonian	91.2%
Galician	91.1%
Catalan	91.0%
Greek	90.9%

Galician (without accents)	90.3%
Portuguese (without accents)	89.6%
Hungarian	89.0%
Icelandic	88.5%
Croatian	87.9%
Czech (without accents)	87.8%
Italian, medieval	87.4%
Italian, medieval (without accents)	87.4%
Romanian	86.5%
Hebrew	86.3%
Croatian (without accents)	86.2%
Ukrainian	85.8%
Italian, modern	85.6%
Finnish	85.4%
Latin, classical	83.9%
Swedish	83.9%
Polish	82.0%
Czech	81.9%
Albanian	81.8%
Welsh	79.6%

language	correlation with Language B
Macedonian	97.7%
Italian, modern (without accents)	97.5%
Italian, modern	97.4%
Italian, medieval	97.3%
Italian, medieval (without accents)	97.2%
Estonian	97.1%
Catalan	96.9%
English	96.9%
Portuguese	96.8%

Galician	96.8%
Spanish	96.6%
Russian	96.6%
Galician (without accents)	96.5%
Spanish (without accents)	96.4%
Croatian	96.2%
Finnish	96.2%
Latin, classical	96.1%
Turkish	95.9%
Romanian	95.9%
Greek	95.8%
Hebrew	95.5%
Hungarian	95.5%
Catalan without accents	95.4%
Icelandic	95.3%
Croatian (without accents)	95.2%
Portuguese (without accents)	94.9%
Swedish	94.4%
Albanian	94.3%
French	94.2%
Latvian	94.0%
Ukrainian	93.7%
Romanian (without accents)	93.7%
German	93.1%
Welsh	92.4%
Polish	92.2%
Irish	92.0%
English, medieval	91.3%
Czech (without accents)	91.0%
Czech	90.2%
Norwegian	90.1%
Danish	90.1%

French (without accents)	90.0%
Dutch	87.9%
Belarusian	76.3%

Sources: author's analysis based on letter frequencies. Frequencies in modern European languages by courtesy of Stefan Trost Media; except in Hungarian, Romanian and medieval Italian by author; glyph frequencies in Languages A and B by author.

medieval or classical language

BACK TO THE "GOLD BUG"

We therefore attempted mappings of selected extracts from the Voynich manuscript, in Currier's Language A and in Language B.

At this point the question arose: what passages to select? It was tempting to use the first page (f001r), which had often been a focus of other research, on the grounds that this page might contain a description of the contents of the manuscript. The last page (f116v) has also been the object of intensive study, but most of its glyphs are indecipherable and some of them might not even be Voynich glyphs at all. It might also be that not all of the manuscript has meaning; there could be pages with meaningful text, and others filled with meaningless gibberish.

On the other hand, we might reasonably expect that the most frequent "words" in the manuscript are meaningful.

We therefore attempted transliterations of the six most common "words" from the v101 alphabet into several selected medieval languages, in which the letter frequencies had high correlations with the Voynich glyph frequencies.

The "Gold Bug" process, for each of Languages A and B, and for each of the selected target languages, was as follows:

- In Language A and Language B separately, we identified the five glyphs with the highest probability of representing vowels, using the Sukhotin algorithm and Dr Mans Hulden's

implementation of the algorithm in the Python programming language. (The Sukhotin algorithm is explained in Appendix 5.)

- In each language, we assumed that all other glyphs represented consonants.
- We hypothesized that the most frequent vowel-glyph, which was o {o} in A and 9 (9) in B, could map to the most frequent vowel in the candidate language, which in most cases was "e."
- We mapped the second most frequent vowel-glyph, 9 {9} in A and c (c) in B, to the second most frequent vowel in the candidate language, which was typically "a," "e," "i," or "o."
- We mapped the third most frequent vowel-glyph, which was a (a) in A, and o {o} in B, to the third most frequent vowel in the candidate language, which was typically "a," "i," or "o."
- We mapped the fourth and fifth most frequent vowel-glyphs in the same way.
- We mapped the consonant-glyphs to consonants in descending order of frequency.

Table 8.2A. Voynich manuscript, Language A, v101 transliteration: mappings of the most common "words" to selected European languages, on the basis of glyph and letter frequencies

Voynich manuscript			languages								
			Middle English, c.1340		Galician-Portuguese, 1220-1300		Middle French, c.1456		Old Albanian, 1555		
		RSQ =>	95.9%		95.8%		94.9%		94.4%		
"word"	key	% of glyphs	string	% of letters	string	% of letters	string	% of letters	string	% of letters	st
ꝺa ıı∀	8am	0.6%	TID	0.0%	NOQ	0.0%	RID	0.0%	HAL	0.0%	N
ɔɔoɬ	1oe	0.6%	NES	0.1%	RES	0.1%	SEN	0.1%	TEN	0.2%	T
ƨ	s	0.5%	L	4.1%	V	2.4%	P	2.7%	P	2.3%	
ɔɔoʔ	1oy	0.4%	NEÞ	0.0%	RED	0.1%	SEL	0.0%	TEM	0.1%	T
a ıı∀	am	0.3%	ID	0.3%	OQ	0.0%	ID	0.1%	AL	0.3%	A

ol	oe	0.3%	ES	0.9%	ES	1.3%	EN	2.3%	EN	2.7%	E

Source: author's analysis, based on mappings in Appendix 1A. RSQ (R squared) refers to coefficient of correlation between glyph frequencies and letter frequencies.

String counts from following corpora: *Auchinleck manuscript* (English); *Cantigas d'Amigo* (Galician-Portuguese); *Meshari* (Albanian); *La Farce de Maistre Pathelin* (French); OVI (Italian); LatinISE medieval subcorpus (Latin).

common word in respective corpus

common string in respective corpus

Table 8.2B. Voynich manuscript, Language B, v101 transliteration: mappings of the most common "words" to selected European languages, on the basis of glyph and letter frequencies

			languages							
Voynich manuscript			Italian, medieval		Latin, medieval		Galician-Portuguese, 1220-1300		Middle English, c.1340	
			96.5%		92.8%		89.8%		82.8%	
"word"	key	% of letters	string	% of letters	string	% of letters	string	% of letters	string	% of letters
aiin	am	0.5%	OD	0.1%	UC	0.1%	IQ	0.0%	AD	0.3%
ol	oe	0.5%	IT	0.5%	AN	0.6%	ON	1.5%	IT	0.4%
ar	ay	0.3%	OS	0.7%	UR	0.4%	ID	0.2%	AÞ	0.1%
cccs9	1c89	0.3%	LANE	0.0%	MITE	0.0%	MARE	0.0%	RONE	0.0%
or	oy	0.3%	IS	0.5%	AR	0.2%	OD	0.3%	IÞ	0.4%
s8aiin	8am	0.2%	NOD	0.0%	TUC	0.0%	RIQ	0.0%	NAD	0.0%

Source: author's analysis, based on mappings in Appendix 1B. RSQ (R squared) refers to coefficient of correlation between glyph frequencies and letter frequencies.

String counts from following corpora: OVI (Italian); LatinISE (Latin); *Cantigas d'Amigo* (Galician-Portuguese); *Auchinleck manuscript* (English); *La Farce de Maistre Pathelin* (French); *Meshari* (Albanian).

common word in respective corpus

common string in respective corpus

Once we had a set of provisional transliterations, we checked the resulting letter strings against corpora of the respective medieval languages, to determine whether any of these strings occurred in those languages; if so, whether they occurred as words, and if so, whether they were common or rare words.

The details of the frequency distributions and the correlations need not concern us here; they are summarised in Appendixes 1.1 through 1.4. We will cut to the chase and summarize, in Tables 8.2A and 8.2B, our first attempts at mapping of the most frequent Voynich "words."

These mappings do indeed produce a few recognizable words in medieval languages. They are not necessarily the words that we would expect.

In the case of Language A, only Old Albanian and medieval Latin yield real words in two out of six cases; and middle French in one case out of six.

In the case of Language B, the mappings yield just two short words: one in Middle English and one in Middle French. However, many of the mapped text strings are common in European languages as parts of words, but not as whole words. We might conjecture that the spaces in the Voynich manuscript do not necessarily correspond to word breaks in the presumed precursor documents.

The reason that so many European languages correlate well with Voynich languages A and B is, we think, that most of these languages are part of the Indo-European family and therefore have similarities in structure, word roots and letter frequencies. The differences in correlations are quite small, of the order of a few percentage points, and do not give us a robust method of prioritizing the candidate languages.

All of the above is as we could, and should, expect. We would be amazingly fortunate if the frequencies of glyphs and letters lined up so well that we could map glyphs to letters on a one-to-one basis. We should remember also that the Voynich manuscript, if it is based on natural languages, is a text (and with about 40,000 "words," not a very a large one) and not a corpus. We have seen that *La Divina Commedia*, which has about 100,000 words, has some divergences in its letter frequencies from the corpus of the medieval Italian language. Likewise, if the Voynich manuscript was derived in some way from a document in a European

language, the frequencies of its glyphs do not have to be the same as the letter frequencies in the broad corpus of the original language.

When William Legrand in "The Gold Bug," and Sherlock Holmes in "The Adventure of the Dancing Men," looked for meaning in the character frequencies, they did not expect to find an exact lineup of the frequencies with those of the English language. They hoped to find, and they did find, some strings of letters which made sense; and we can presume that as often as not, they filled in the gaps by guesswork. Our task is more difficult than theirs because our transliterations so far have not yielded many words which invite guesswork or swapping pairs of letters.

We must also admit that we have tested only six corpora of languages that were written and spoken in medieval times—Old Albanian, Middle English, Old French, Galician-Portuguese, Italian, and Latin. If we were to undertake this task with the diligence that it deserves, we would need to explore corpora of a wider spectrum of medieval European languages. That is a task beyond the scope of this book.

ALTERNATIVE KEYBOARD ASSIGNMENTS

In any case, before we take the process of frequency analysis too far, we need to recall that up to now, we have been working not from the original glyphs, but from the v101 keyboard assignments. In Chapter 3, we expressed various reservations or apprehensions about these assignments. We may recapitulate our thoughts as follows:

- There are instances where several distinct keyboard assignments could be reconstrued as representing a single glyph, with nothing more than variations in handwriting.
- There are instances where a keyboard assignment, which the v101 alphabet represents as a single glyph, could be reconstrued as representing several glyphs, written cursively.

We therefore constructed several alternative sets of keyboard assignments, which with due respect to Glen Claston, we named v102, v103, and v104 (not wishing to exclude that there might subsequently be a v105 or higher).

One way to construct such an alphabet would be simply to look at the visual resemblances between glyphs: for example, among the most-common glyphs, between {8} and {7}, and between {m} and {n}. We could assume that the minor differences are simply variations in handwriting, or even errors; and on that assumption, we could combine groups of glyphs that seem similar.

However, there is a better way. If we suspect two glyphs to be the same, we could identify all the Voynich "words" containing one of them (say, the less common one); and replace every occurrence of this glyph with the more common glyph. If the two glyphs are the same, then after the replacement we should still have a Voynich "word"; it should occur elsewhere in the manuscript, and we can count those occurrences.

In Table 8.3, we have summarized the results of this test, for selected glyphs and glyph strings that we suspected to be duplicates; and we included a comparison with modern American English, to show to what extent common modern vowels and consonants can be interchanged and still make words.

Table 8.3. Voynich manuscript, v101 transliteration: effect of replacing selected glyphs or glyph strings; and comparison with modern American English

selected string			number of "words" containing string	number of "words" tested	replace-ment string		words with string replaced			
glyph	v101	position			glyph	v101	count	percent	word	
Voynich manuscript										
	a	any	3,466	100		o	74	74%		
	7	any	951	100		8	99	99%		
	m	any	624	100		n	76	76%		
	cc	any	686	100		C	93	93%		
	A	any	460	100		a	74	74%		

◌	A	any	460	100	◌	o	92	92%	◌	
◌	n	any	337	100	◌	m	73	73%	◌	
◌	3	any	238	50	◌	2	49	98%	◌	
◌	5	any	212	50	◌	2	43	86%	◌	
◌	6	any	83	20	◌	8	17	85%	◌	
◌	N	any	74	30	◌	m	25	83%	◌	

modern English (Brown University corpus)

e	any	29,630	100	a	9	9%	then	
e	any	29,630	100	o	6	6%	been	
e	any	29,630	100	i	5	5%	then	
t	any	19,980	100	n	8	8%	to	2
t	any	19,980	100	s	15	15%	the	6
t	any	19,980	100	h	7	7%	still	

Source: author's analysis. In this table, "word" denotes a string of glyphs delimited by spaces, uncertain spaces or line breaks; or an American English word appearing in the Brown University corpus.

replaceable in 50% percent of cases or more

replaceable in less than 50 percent of cases

These results seem to tell us the following:

- The interchangeability of the selected glyphs is vastly greater than that of the most-common vowels in English, or of the most-common consonants in English.
- The glyphs ◌ {m}, ◌ {n} and ◌ {N} are clearly related, in more than just a visual sense; they have a common vocabulary. We could imagine that ◌ {n} and ◌ {N} are incomplete versions of ◌ {m}, or that ◌ {m} and ◌ {n} are embellished versions of ◌ {N}; even that the Voynich scribes were free to interchange these glyphs according to their predilections.

- The glyphs ȣ {8}, ȣ {7}, ȣ {6}, and ȣ {&} are probably the same.
- The glyphs ᴄ̂ {2}, ᴄ̊ {3}, ᴄ̂ {5}, ᴄ̌ {!}, ᴄ̈ {#}, and ᴄ̊ {+} are probably the same; and all of them may be accented versions of ᴄ {1}.
- The glyph pair ᴄᴄ {cc} is probably the same as ᴄ {C}.
- The glyph ɑ {A} is possibly the same as ᴏ {o}; there is even a suspicion that ɑ {a} is the same as ᴏ {o}.

Table 8.4. The most frequent glyphs in the v101 alphabet, and some alternatives (v102, v103, v104)

glyph	note	v101		v102		v103		v104		v104 vs. v101
		key	count	key	count	key	count	key	count	
ᴏ	incl {4o}	o	25,043	-	-	-	-	-	-	-
ᴏ	excl {4o}	o	19,778	o	20,546	o	20,546	o	20,546	↓
ꝯ		9	17,436	9	17,838	9	17,838	Φ	17,838	
ɑ		a	14,496	a	14,496	a	14,496	a	14,496	
ᴄ		c	13,750	c	14,834	c	14,915	c	14,935	↑
ᴄᴄ		1	11,052	1	11,052	1	17,833	1	17,833	↑
ɾ		e	10,653	e	10,653	e	10,653	e	10,653	
ȣ		8	10,343	8	13,208	8	13,208	8	13,208	↑
ℍ		h	9,919	h	9,919	h	10,831	h	10,946	↑
ʔ		y	6,717	y	7,548	y	7,548	Ψ	7,548	↑
ℍ		k	6,118	k	6,118	k	7,093	k	7,093	↑
↾	incl {4o}	4	5,408	4	-	-	-	-	-	-
↾ᴏ	single glyph	4o	5,265	④	5,265	④	5,265	Δ	5,265	↓
ıı℧		m	4,112	îN	-	iiN	-	îΩ	-	-
ᴄ̂		2	3,359	2	4,603	1’	-	1’	-	-
ᴄᴄ		C	2,844	C	2,844	cc	2,844	ĉ	2,877	
ȣ		7	2,714	8	-	8	-	8	-	-

꒐		s	2,702	s	2,836	s	2,836	s	2,836	
꒑		n	1,770	iN	-	iN	-	iΩ	-	-
꒘		p	989	p	1,229	p	1,229	Π	1,229	↑
꒦		K	910	K	910	k1	-	k1	-	-
꒻		g	879	g	879	g	1,665	g	1,665	↑
꒦		H	866	H	866	h1	-	h1	-	-
꒡		A	768	o	-	o	-	o	-	-
꒻		j	564	j	564	g	-	g	-	-

Source: v101 assignments from Zandbergen, Rene: *Text Analysis - Transliteration of the Text*, Table 6: the v101 basic character set; v102, v103 and v104 assignments by author.

change from v101 keyboard assignments

"marker glyph" in initial position

"marker glyph" in final position

We present the the essence of our proposed v102, v103, and v104 assignments in Table 8.4. These alphabets have three main innovations:

- We combine those glyphs that, from Table 8.3, appear to be the same, as well as some variants which are relatively infrequent.
- However, in case there is some meaning attached to the ꒑ shape embedded in ꒑꒑ {m} and ꒑ {n}, we retain this shape and represent it by the v101 assignment {i}. If in due course it appears in our transliterations as a redundant letter, we can always discard it.
- Similarly, we disaggregate the ꒑ shapes that we see embedded in glyphs such as ꒑꒑ {I} and ꒑꒘ {P}.
- We have to consider the possibility that the eight "pedestal glyphs," namely the "gallows" glyphs which are superimposed on a lengthened ꒾ {1}, which looks like a pedestal, are not distinct glyphs. For example, the glyph ꒦ {K} could be a ꒻

{k} plus a lengthened {1}, and therefore, not knowing the intended order of these two glyphs, we could give it the keyboard assignment {k1} or {1k}. If we get the order wrong, or if the lengthened is not the same as a normal {1}, it may become obvious in the eventual transliteration, and we can change the keyboard assignments.

- We also have four cases of "grand piano glyphs" (that is, "gallows" glyphs superimposed on a). This is a shape for which there is no keyboard assignment in v101, but since it resembles a lengthened {1} attached to a {c}, we could provisionally assign it as {1-c}. An example is {L}, which we could think of as a {k}, plus a {1}, and a {c}, and to which we could assign {k1c} on the keyboard.

The v104 alphabet has only twenty-four glyphs that are intended to be transliterated. Here we draw upon Mary D'Imperio's research, which we reported in Chapter 7, on the very specific physical positioning of certain glyphs within the "word." This led to the idea that at least eleven glyphs are "marker glyphs" which represent something like punctuation or abbreviations in prefixed or suffixed form, but do not map to any specific letter in a natural language.[4] They are practically always initial or final glyphs.

Our idea was that in any transliteration, the Greek letters for the "marker" glyphs will stay as they are, to be figured out later. For example, if we were lucky enough to produce a Latinate string such as MAXIMΦ, we might reasonably think that our transliteration was sound and that the Φ represented the Latin suffix -US or -UM.

We attempted one final set of transliterations, using the v104 alphabet. The "marker" glyph [Φ] was excluded from the frequency calculations and therefore from the transliterations. The remaining glyphs were designated as "normal" glyphs.

Table 8.5A. Voynich manuscript, Language A, v104 transliteration with Φ removed: mappings of the most common "words" to selected European languages, on the basis of glyph and letter frequencies

Voynich manuscript			European languages					
glyphs	key	word count	Galician-Portuguese, 1220-1300	Middle French, c.1456	Old Albanian, 1555	Latin, medieval	Middle English, c.1340	Italian mediev
		R squared =>	92.9%	90.4%	90.0%	89.2%	88.2%	77.4%
ꞔɑ ıı\	8aîΩ	482	SOLI	NIMO	NARU	SADU	SIHA	RICO
ƈƈoᵩ	1oe	274	AEN	AER	IEH	IEN	OET	AET
ƨ	s	238	Q	D	L	C	D	D
ƈƈoʔ	1oΨ	185	AEM	AET	IEſ	IEM	OER	AEL
ɑ ıı\	aîΩ	140	OLI	IMO	ARU	ADU	IHA	ICO
oᵩ	oe	139	EN	ER	EH	EN	ET	ET

Source: author's analysis, derived from proposed mappings in Appendix 1C (excluding Φ).

common word in respective corpus

common string in respective corpus

rare word in respective corpus

Corpora of reference: *Meshari* (Albanian); Auchinleck manuscript (English); *La Farce de Maistre Pathelin* (French); *Cantigas d'Amigo* (Galician); OVI (Italian); LatinISE medieval subcorpus (Latin).

Table 8.5B. Voynich manuscript, Language B, v104 transliteration with Φ removed: mappings of the most common "words" to selected European languages, on the basis of glyph and letter frequencies

Voynich manuscript			languages					
"word"	key	word count	Italian, medieval	Latin, medieval	Galician-Portuguese, 1220-1300	Middle English, c.1340	Middle French, c.1456	Old Albanian, 1555
		R squared =>	93.9%	89.6%	86.7%	77.3%	72.6%	61.1%
oᵩ	oe	684	ET	EN	EN	ET	ER	EH
ɑ ıı\	aîΩ	620	ODU	UCO	IQU	ADU	ODU	ULO

⟃⟑ƈ8	1c8	602	IAN	AIT	OAR	ION	IAS	AIT
⟃ʔ	aψ	510	OL	UM	IM	AR	OT	Uſ
⟃̓⟑ƈ8	1’c8	493	I’AN	A’IT	O’AR	I’ON	I’AS	A’IT
oʔ	oψ	395	EL	EM	EM	ER	ET	Eſ

Source: author’s analysis, based on proposed mappings in Appendix 1D, excluding Φ.

Note: the accent on ⟃̓ and its variants was not mapped.

common word in respective corpus

common string in respective corpus

rare word in respective corpus

Corpora of reference: *Meshari* (Albanian); Auchinleck manuscript (English); *La Farce de Maistre Pathelin* (French); *Cantigas d’Amigo* (Galician); OVI (Italian); LatinISE medieval subcorpus (Latin).

Here we thought we saw a significant improvement in the transliterations. In most of the candidate medieval languages that we tested, the mappings produced a few short words, or common strings, that could be found in the respective corpora.

A more intensive examination of “marker” glyphs, and a wider selection of target languages, may be productive areas of investigation.

OTHER PATHWAYS

At this point in our analysis, it seemed that we could make the following statements of probability:

- If there existed a one-to-one mapping between some of the Voynich glyphs and the letters in European languages, then with a high probability, the ubiquitous o {o} mapped to E; the extremely common 9 {9}, if it mapped to any letter, probably

mapped to A; and ᴄ {c} and ᴀ {a} probably mapped to common vowels.

- It was very probable that the specific string ᔭo {4o} was a single glyph.
- There was a significant probability that ᔭo {4o}, ᔭ {4}, all the glyphs that contained ᑎ {N}, ᕐ {y} or its variants, or ᘐ {*} or its variants, and even ᓐ {9}, did not map to any specific letters in European languages, but instead were what we have called "marker glyphs," which signified "beginning of word" or "end of word" or certain (possibly variable) prefixes or suffixes.
- It was even possible that the "marker glyphs" had no meaning at all and were simply what Prescott Currier might have called "junk filler." In other words, if there was any meaning in the Voynich manuscript, it might reside in what remained after the "marker glyphs" had been removed or neutralized.
- It was even conceivable that the ᴏ {o}, even though it often occurred internally to a "word," might also be a "marker glyph" with some as yet unknown function or with no meaning. In that case we, or the interested reader, would need to construct a v105 alphabet in which the {o} was also neutralized along with the other "marker glyphs."

Nevertheless, after all this juggling with alphabets, we had to consider the possibility that the "Gold Bug" approach would not work. We began to surmise that, if the manuscript was based on natural European languages, its relationship to those languages was not as straightforward as a one-to-one mapping between glyphs and letters.

We could think of several possibilities:

- It was common in medieval Latin and Italian manuscripts to abbreviate the text, often because the parchment was expensive, and abbreviation saved space. The Voynich manuscript does not seem to have been short of space, because there are many luxurious blank areas which could have been filled with glyphs.

But perhaps the original document had pervasive and even radical abbreviations, and the scribes simply preserved them. It would have been logical for the project manager to give such an instruction to the scribes; otherwise, the scribes would have had the time-consuming and error-prone task of expanding the abbreviations. In this scenario, the original letter frequencies would have been distorted and the transliteration to Voynich would have preserved the distortion.

- The Voynich text may have been enciphered, in which case our transliterations represent ciphertext rather than plain text and await a further stage of decipherment.
- Or the Voynich text may have been based on languages other than the thirty-seven that we have considered.

GLYPH POSITIONS

Before we abandon the "Gold Bug" approach, it is perhaps worthwhile recalling an idea that emerged from the work of Mary D'Imperio, which we narrated in Chapter 6. D'Imperio (if we may be permitted to interpret her conclusions), did not think that the glyphs could have been mapped one to one from letters in a natural language. There was simply too much positioning among the glyphs; there were distinct species of glyphs, such as initial glyphs, internal glyphs, and final glyphs.

D'Imperio identified at least six or seven distinct classes of position for a glyph within a Voynich "word." For our testing, given the functions available in Microsoft Word and Excel, we had to be content with four: isolated (meaning a single-glyph "word"), and the initial, internal, and final positions in a "word" of at least two glyphs.

We would have liked to test this hypothesis with corpora of several different medieval languages; the most-accessible databases were the OVI corpus of medieval Italian, the medieval subcorpus of the LatinISE corpus, and the epic romance *Partonopeus de Blois*, of unknown authorship, serving as a corpus for Old French.

Our methodology for the testing was as follows. We assumed that:

- The Voynich scribes started with documents in medieval Italian, Latin, or French, which they had to transcribe to glyphs.
- They mapped spaces in the source documents to spaces in their manuscript.
- For the single-letter words in the source language, such as E and A in Italian, they had corresponding glyphs.
- For all words of two letters in the source language, they had a set of glyphs corresponding to the initial letters, and another (overlapping) set corresponding to the final letters. Thus, a given glyph could have one meaning as an initial glyph and another as a final letter.
- For all words of three letters or more in the source language, they had a set of glyphs corresponding to the initial letters, another corresponding to the internal letters, and another set corresponding to the final letters. Again, these three sets could overlap.
- Using the available corpora, we calculated the frequencies of letters in each of the four positions relative to words.
- Using the portion of the Voynich manuscript that was in Currier's Language B, and applying our v104 alphabet, which seemed the most appropriate for this purpose, we calculated the frequencies of the glyphs in each of the four positions, relative to Voynich "words."
- We lined up the rankings of letters and glyphs, with a view to establishing an approximate correspondence in each of the four positions.

The results are summarized in Table 8.6.

So far, so good. The correlations were quite respectable, although none of the three languages was consistently a better map than the others.

The results supported what Mary D'Imperio suspected: that it was possible for the meaning (or transliteration) of a glyph to vary depending on where the glyph was positioned within the string. More accurately, we should say that we could find statistically robust mappings on the basis of

this hypothesis. For example, if we wanted to map from glyphs to medieval Italian, the glyph ο {o} if occurring alone could be the Italian word I ("the"); if an initial glyph, it could represent a D; if an internal glyph, it could be a T; and if a final glyph, it could be an N.

This hypothesis mitigated the confusion to which we referred in Chapter 3, regarding a text which seemed to have at least two glyphs, namely, ο {o} and ϙ {9}, that were far too frequent to map to a single letter in a natural language.

Table 8.6. Voynich manuscript, Language B: comparisons of position-dependent glyph frequencies and position-dependent letter frequencies in medieval Italian and Latin

single-letter words

Voynich Language B			medieval Italian		medieval Latin	
glyph	v104	%	word	%	word	%
ϧ	e	19.0%	E	57.2%	A	72.6%
ϙ	Φ	17.1%	A	21.2%	O	5.1%
ϩ	Ψ	15.5%	È	11.1%	E	5.0%
ϲ	s	13.9%	I	4.5%	B	3.5%
ο	o	9.9%	O	4.4%	C	3.2%
†ο	Δ	9.6%	À	0.9%	D	1.7%
ϑ	8	4.5%	U	0.3%	I	1.6%
ⲥ	1	1.6%	Ò	0.3%	M	1.3%
ⲥ̀	2	1.6%	D	0.0%	N	1.1%
⍺	a	1.4%	Y	0.0%	S	1.1%
			RSQ	52.6%	RSQ	30.1%

initial letter in words of two or more letters

Voynich Language B			medieval Italian		medieval Latin	
glyph	v104	%	letter	%	letter	%
ο	o	23.1%	D	11.6%	E	14.1%
†ο	Δ	17.5%	C	11.1%	S	9.5%
ⲥ	1	16.0%	S	8.7%	I	8.5%

ɑ	a	9.5%	E	8.4%	Q	8.2%
☍	2	9.3%	A	8.0%	P	7.4%
ȣ	8	7.7%	L	7.9%	A	7.2%
ɾ	e	5.1%	P	7.8%	D	6.3%
ᚻ	h	3.8%	I	4.8%	N	5.5%
ዐ	Φ	3.7%	M	4.6%	C	5.1%
ᚠ	k	2.8%	F	3.9%	U	4.0%
			RSQ	82.8%	RSQ	89.9%

internal letter in words of two or more letters

Voynich Language B			medieval Italian		medieval Latin	
glyph	v104	%	letter	%	letter	%
ɔ	c	14.8%	E	13.1%	U	13.7%
ɑ	a	11.4%	R	9.6%	I	12.0%
ȣ	8	10.2%	N	8.5%	E	9.7%
ᚻ	h	8.8%	A	8.3%	O	7.4%
ο	o	5.3%	I	8.1%	A	6.9%
ᚠ	k	4.5%	T	7.7%	N	6.8%
☌	1	4.1%	O	7.4%	T	6.7%
ᛁᛁ	î	4.1%	L	5.8%	R	6.4%
ᛁ	i	3.9%	U	5.5%	S	4.8%
☾	ĉ	3.5%	S	4.8%	L	4.6%
			RSQ	82.4%	RSQ	90.7%

final letter in words of two or more letters

Voynich Language B			medieval Italian		medieval Latin	
glyph	v104	%	letter	%	letter	%
ዐ	Φ	46.1%	E	24.2%	T	17.0%
ᘐ	Ω	17.7%	O	21.3%	M	16.7%
ɾ	e	15.4%	A	18.1%	S	13.6%
ᛇ	Ψ	15.1%	I	16.6%	E	11.8%
ο	o	2.8%	N	4.5%	A	6.8%

ϱ	Π	2.7%	L	3.8%	O	6.2%
ƨ	s	2.1%	R	3.0%	I	6.0%
ϑ	8	1.9%	U	1.2%	N	6.0%
⁋	h	0.4%	Ì	1.2%	D	5.2%
ɑ	a	0.2%	Ò	1.2%	R	3.8%
			RSQ	78.1%	RSQ	75.7%

Note: Counts in Italian based on OVI corpus; counts in Latin based on LatinISE medieval subcorpus.

Note: Correlations are based on the ten most frequent letters and glyphs in each position.

Table 8.7. The Voynich manuscript, Language B: the ten most frequent "words", and possible mappings to medieval Italian based on positional frequencies

		Voynich manuscript, v104 transliteration				Italian
"word"	key	initial glyph	internal glyphs	final glyph	% of words	string
oϱ	oe	o	-	e	2.3%	DA
ɑıı∀	aîΩ	a	î	Ω	2.2%	ELO
ɔɔcϑϱ	1c8Φ	1	c8	Φ	2.2%	SENE
ć2cϑϱ	1'c8Φ	1	'c8	Φ	1.8%	S?ENE
ɑʔ	aΨ	a	-	Ψ	1.6%	EI
ϑɑıı∀	8aîΩ	8	aî	Ω	1.4%	LRLO
oʔ	oΨ	o	-	Ψ	1.3%	DI
ɑϱ	ae	a	-	e	1.2%	EA
ɔɔcϱ	1cΦ	1	c	Φ	1.1%	SEE
†o⁋ɑı∀	ΔhaiΩ	Δ	hai	Ω	1.1%	CARUO

Source: author's analysis, based on positional frequency comparisons in Table 8.6

Note: the accent over the glyph ć and its variants is not mapped.

common word in OVI corpus

rare word in OVI corpus

However, our investigation of this hypothesis was far from complete. If these mappings had any validity, we should be able to produce

transliterations of some passages of the Voynich manuscript, and some portions of these transliterations should make sense.

As an alternative to selecting a random line from the manuscript, we identified the ten most common "words" in Language B, applied the mappings from Table 8.6, and ran searches on several corpora of medieval languages to see whether we had produced real words, and if so, how common these words were in the respective corpora. The results are shown in Table 8.7.

The mappings to medieval Italian were encouraging. We had produced two extremely common Italian words: DA ("from"), and DI ("of").

There were also at least two promising avenues for refinement of the mapping. The Italian strings SENE and S?ENE look as if they should read BENE ("well"); in which case an initial {1} or {2} or both, might map to an Italian initial B. The mapping of {am} to the Italian word ELO would yield a much-higher count if ELO were replaced by LO ("the," masculine singular); in that case a final {m} (and it is nearly always final) might map simply to an Italian final O. That in turn would offer alternative mappings for the string {8am}, which on our first mapping did not yield an Italian word; the possibilities include ALO or ALLO ("at the" or "to the").

On the other hand, the mappings to medieval Latin and to Old French produced no results of any consequence.

In this light of this rudimentary test, the concept of position-dependent mappings between glyphs and letters seems a promising area for research. There is clearly also scope for applying the concept to a mapping between the Voynich manuscript and medieval languages that we have not yet tested.

GLYPH PAIRS

Mary D'Imperio, in her technical papers, also raised the hypothesis that, if the glyphs mapped in some ways to natural letters, the mapping was not necessarily based on single glyphs, but possibly on strings of glyphs of variable length.

The simplest (though not the only) interpretation of this hypothesis is that glyph pairs were the building blocks of the manuscript. If that were the case, the left-hand glyph of a pair would naturally show up frequently as an initial glyph, and the right-hand glyph would often be a final glyph. We realized that we might also need to consider glyph triplets, and more.

Now it was time to put this hypothesis to the test.

It is quite a straightforward exercise to count all the most frequent glyph pairs in the Voynich manuscript. The "find and replace" function in Microsoft Word is up to that job. Having done that, we could compare the frequencies of the glyph pairs with the frequencies of letter pairs (digraphs) in some European languages. We could then, as we did earlier for single glyph-to-letter mappings, look at the correlations and see which languages are better matches for the Voynich manuscript.

However, we need to recall (from the work of Currier and D'Imperio) that the glyphs are not distributed uniformly within the "words." There are some glyphs that appear predominantly at the beginnings of "words," and others that are predominantly at the end. If this is true of glyphs, it must also be true of glyph pairs. For example, the glyph pair {1c} is very common as the beginning of a Voynich "word"; the digraph RE is very common as the end of an Italian word—as in the song "Volare." However, even if the frequencies were identical, we would not expect {1c} to map to RE, because the positioning is wrong.

We therefore counted the position-dependent numbers of glyph pairs in the Voynich manuscript (focusing on Language B), ranked them in order of frequency, and matched them with the frequencies of digraphs in medieval Italian (from the OVI corpus), classical Latin (from the Packhum corpus), and Old French (from *Partonopeus de Blois*). The most-interesting results are summarized in Table 8.8.

Table 8.8. Voynich manuscript, Language B, v104 transliteration: position-dependent counts of the most common glyph pairs, by position, and the most common digrams in medieval Italian

isolated position (two-letter words)					
Voynich manuscript, Language B				medieval Italian	
glyph pair	v101	v104	frequency	digram	frequency

ol	oe	oe	0.54%	DI	0.66%
ar	ay	aΨ	0.40%	LA	0.57%
or	oy	oΨ	0.30%	IN	0.37%
al	ae	ae	0.29%	DE	0.24%
89	89	8Φ	0.19%	LO	0.24%
4ol	4oe	Δe	0.15%	IL	0.22%
19	19	1Φ	0.05%	SI	0.20%
1o	1o	1o	0.04%	LI	0.20%
29	29	2Φ	0.04%	LE	0.19%
4oh	4oh	Δh	0.03%	SE	0.15%
				R squared	88.6%

initial position

Voynich manuscript, Language B				medieval Italian	
glyphs	v101	v104	frequency	digram	frequency
4oh	4oh	Δh	2.40%	CO	1.07%
1c	1c	1c	1.91%	CH	1.06%
ok	ok	ok	1.42%	QU	0.84%
oh	oh	oh	1.37%	DE	0.79%
2c	2c	2c	1.31%	PE	0.76%
8a	8a	8a	1.10%	DI	0.53%
4ok	4ok	Δk	0.81%	NO	0.50%
oe	oe	oe	0.71%	AL	0.40%
1o	1o	1o	0.40%	PA	0.34%
eh	eh	eh	0.39%	MA	0.32%
				R squared	95.7%

internal position

Voynich manuscript, Language B				medieval Italian	
glyphs	v101	v104	frequency	digram	frequency
c8	c8	c8	3.82%	ON	1.60%
ha	ha	ha	2.10%	ER	1.39%

ſſc	hc	hc	1.45%	EN	1.30%
ctc	1c	1c	1.41%	AN	1.22%
cc	cc	cc	1.13%	NT	1.03%
ſſcc	hC	hĉ	1.04%	LL	0.98%
ſſa	ka	ka	1.03%	IA	0.94%
co	co	co	1.00%	AR	0.91%
8a	8a	8a	1.00%	OR	0.88%
ſſc	kc	kc	0.79%	RO	0.78%
				R squared	81.8%

final position

Voynich manuscript, Language B				medieval Italian	
glyphs	v101	v104	frequency	digram	frequency
89	89	8Φ	5.27%	TO	1.08%
IN	IN	îΩ	2.47%	RE	1.04%
c9	c9	cΦ	1.79%	NO	0.90%
ay	ay	aΨ	1.65%	HE	0.78%
iN	iN	iΩ	1.40%	TE	0.71%
ae	ae	ae	1.22%	LI	0.54%
oe	oe	oe	1.06%	RA	0.53%
C9	C9	ĉΦ	0.86%	ER	0.53%
oy	oy	oΨ	0.54%	TA	0.50%
p	p	iΠ	0.51%	TI	0.49%
				R squared	69.1%

Source: author's analysis.

Note: The table shows only the ten most frequent glyph pairs and digraphs in each position. R squared refers to the correlations based on the 62 most frequent glyph pairs and digraphs in each position. Frequencies in the Voynich manuscript are expressed relative to the number of glyphs in the v104 transliteration. Frequencies in medieval Italian are expressed relative to the number of letters in the OVI corpus.

Of the three languages, medieval Italian gave the best correlation (95.7 percent) for the initial glyph pairs; Latin and Old French (which are not shown in the table) had much lower correlations, in the range 80 to 86

percent. This is as we would expect from the distance-decay hypothesis, Old French being geographically more distant from Frascati, and Latin being more remote in time. These results encouraged us to think that medieval Italian could be the underlying stratum at least of Language B.

The issue that now arose was: could we attempt a transliteration from the Voynich manuscript, on the basis of glyph pairs rather than single glyphs?

The obstacle is that we did not know how to parse the glyph strings; and many of the glyph strings have odd numbers of glyphs. For example, one of the most common glyph strings in the manuscript is ⟃⟑ⴱ⟒ [1cΦ]; but this contains two glyph pairs, ⟃⟑ [1c] and ⴱ⟒ [cΦ]. Which one should we map to our target language? Whichever one we chose, there would be one glyph left over for which our correlations provided no guide.

However, we could at least try a transliteration of some common strings of four glyphs, which we could parse into the first pair and the second pair. We tried this exercise with medieval Italian, treating our Table 8.8 as a potential mapping. For what it is worth, for the three most common four-glyph strings in Language B, the exercise yielded one interesting result:

> ⸸ⴰ ⟓⟓⟘ [8aîΩ] ⇔ ⸸ⴰ [8a] + ⟓⟓⟘ [îΩ] ⇔ DI + RE ⇔ DIRE
> which occurs 22,343 times in the OVI corpus and is the medieval (and modern) Italian equivalent of the English “to say.”

Are we onto something here? It could be that our v104 alphabet defined ⸸ⴰ ⟓⟓⟘ [8aîΩ] correctly as a four-glyph “word” (and not three glyphs as in the v101 alphabet). Equally, it could be that ⟃⟑ⴱ⸸⟒ [1c8Φ] and ⟃̓⟑ⴱ⸸⟒ [2c8Φ] are not four-glyph but five-glyph or longer “words.” For example, ⟃⟑ [1] could be ⟃ [é] plus ⴱ [c]; and ⟃̓⟑ [2] could be ⟃ [é] plus ⟒ [Φ] plus ⴱ [c]; in which case, some of our transliterations are producing words of the wrong length, and we need to invent a v105 alphabet.

With that observation, we continued this exercise for six-glyph strings, for which the obvious parsing was first pair + second pair + third pair. For the six most common six-glyph strings in Language B, the most interesting result was as follows:

ofcccd9 [ok1c8Φ]⇔ of [ok] + ccc [1c] + d9 [8Φ] ⇔ QUANTO

Here we had a real word: QUANTO, which occurs 23,408 times in the OVI corpus and is the Italian word for “how much” or “how many.”

To pursue the idea of a mapping between glyph pairs and digraphs would require imagining the mechanism by which the scribes of the Voynich mapped words with odd numbers of letters, including single-letter words.

GLYPH TRIPLETS

There is no reason why we should stop at glyph pairs. We may need also to think about glyph triplets, such as the immensely common cd9 [c8Φ], as possible building blocks of the mapping. Not only that, but we may need to consider that some combinations of glyph triplets, glyph pairs and single glyphs are the building blocks of the manuscript. Mary D’Imperio, in her groundbreaking research in the 1970s, said explicitly that the blocks were probably glyph strings of variable length.

Let us imagine the inaugural meeting of the scribes of the Voynich manuscript, back in the fifteenth century or some later era. The author or producer of the manuscript is giving them a briefing. He asks them to think of themselves as masons, and to imagine the manuscript as a wall that they must build. They have bricks of three sizes: 4-inch, 8 inch, and 12 inch. They should use the 12-inch bricks first, then the 8 inch, if need be, and finally the 4-inch to fill any gaps where the bigger bricks will not go.

The producer then hands out the documents that they must transcribe. They are written, we imagine, in medieval Italian, Latin, or Greek, or some other language that was familiar to professional scribes in Europe in the fifteenth century. We like the idea that these documents are poems, or Christian missals, or some other kind of structured document arranged in lines, with pervasive repetitions.

Finally, there is a distribution of the transcription charts: between strings of letters in the Latin or Greek script, and strings of glyphs in what will be

called, centuries later, the Voynich alphabet. The glyph strings, we imagine, are arranged in columns: one of single glyphs, one of glyph pairs, one of glyph triplets, and maybe there are more. The producer reminds the scribes that, by analogy with masonry, they should use the triplets by preference, then the pairs, then the single glyphs.

This scenario will work only if the scribes map lines in the source documents to lines in the manuscript; spaces in the source documents to spaces in the manuscript; and letter strings in the source documents to glyph strings of the same length. However, no source letter necessarily maps consistently to a unique glyph; the mapping is at the level of strings, and individual glyphs within the strings have no specific meaning.

The scribes must have clear instructions about parsing the words in the source documents. There will be no stopping in midstream and referring to the producer; the fluency of the final document is testimony to that. We can imagine the producer giving instructions along the following lines:

- A single-letter word maps to one (or possibly more than one) of the single glyphs in the reference chart.
- A two-letter word maps to a unique glyph pair in the reference chart.
- A three-letter word maps to a unique glyph triplet in the reference chart.
- A four-letter word is parsed into the first three letters, plus the fourth; the scribes should map them separately using the rules above, but with no intervening space.
- A five-letter word is parsed into the first three letters, plus the last two. And so on … with probably some kind of fallback instruction for letter sequences that are unusual and have not been listed on the chart; in which case the scribes possibly can use letter-by-letter mappings.

Without knowing the language of the source documents, we could imagine that reference charts containing up to fifty single glyphs, fifty glyph pairs and fifty glyph triplets would have been adequate for the scribes to get to work without detailed supervision.

We submitted this hypothesis to the test by investigating to what extent the Voynich manuscript can be constructed from its most-common glyphs, glyph pairs and glyph triplets. The process was as follows (using for simplicity the v101 transliteration):

- identifying the most-common nonoverlapping glyph triplets
- identifying the most-common glyph pairs which were not part of the previously identified triplets
- identifying the most-common glyphs which occurred as single-glyph "words"
- for each glyph string, calculating the total number of glyph occurrences (for example, the three-glyph string c89 [c8Φ] occurs 2,294 times, equivalent to 6,882 glyphs)
- identifying the minimum number of glyph strings in each group, for these strings to account for 85 percent of the Voynich manuscript.

Our results are illustrated in Table 8.9. In summary we found that:

- forty-eight three-glyph strings accounted for 60.4 percent of the manuscript
- thirty-two two-glyph strings accounted for 26.0 percent of the manuscript
- twenty single glyphs (used only as single glyph "words") accounted for 1.3 percent of the manuscript.

Table 8.9. Voynich manuscript, v101/v104 transliteration, the possible building blocks: a minimal set of mutually exclusive glyph strings accounting for at least 85 percent of the text

glyph string	v101	v104	string count	v101 glyph count				
				isolated	initial	internal	final	total
3-glyph strings (48)								
†oll	4oh	Δh	3,022	35	2,975	12	0	9,066
c89	c89	c8Φ	2,294	4	0	17	2,273	6,882

ol9a	oha	oha	1,859	3	797	1,058	1	5,577
ol9c	ohc	ohc	1,564	0	670	892	2	4,692
8a11	8am	8aîΩ	1,148	739	9	0	400	3,444
4o99	4ok	Δk	1,143	25	1,113	5	0	3,429
o9da	oka	oka	1,129	2	751	375	1	3,387
1c8	1c8	1c8	1,105	21	491	555	38	3,315
1co	1co	1co	1,048	126	582	265	75	3,144
c8o	c79	c8Φ	1,034	2	0	12	1,020	3,102
others			16,634	2,449	3,458	4,130	6,597	49,902
total 3-glyph strings			31,980	3,406	10,846	7321	10,407	95,940
							% of manuscript	60.4%
2-glyph strings (32)			20,669					
oy	oy	oy	2,457	440	166	228	1,623	4,914
2c	2c	2c	1,645	17	1,212	409	7	3,290
C9	ĉΦ	ĉΦ	1,215	5	4	6	1,200	2,430
19	1Φ	1Φ	1,002	185	34	22	761	2,004
eh	eh	eh	971	1	404	563	3	1,942
2o	2o	2o	765	101	497	124	43	1,530
ko	ko	ko	745	6	205	496	38	1,490
h9	hΦ	hΦ	739	31	4	28	676	1,478
ho	ho	ho	733	7	161	531	34	1,466
9h	Φh	Φh	702	10	582	108	2	1,404
others			9,695	177	3,315	4,167	2,036	19,390
total 2-glyph strings				980	6,584	6,682	6,423	41,338
							% of manuscript	26.0%
single glyphs (20)			-					
s	s	s	-	449	-	-	-	449
9	9	Φ	-	428	-	-	-	428
e	e	e	-	285	-	-	-	285

۲	y	Ψ	-	284	-	-	-	284
০	o	o	-	264	-	-	-	264
৪	8	8	-	132	-	-	-	132
८	1	1	-	40	-	-	-	40
ℍ	h	h	-	31	-	-	-	31
ɑ	a	a	-	31	-	-	-	31
ℍ	k	k	-	30	-	-	-	30
others			-	128	-	-	-	128
total, single glyphs				2,102	-	-	-	2,102
							% of manuscript	1.3%
Total, above strings (100)			52,649	6,488	17,430	14,003	16,830	139,380
							% of manuscript	87.7%
Total, manuscript								158,963

Source: author's analysis

isolated strings occurring over 100 times

initial strings occurring over 100 times

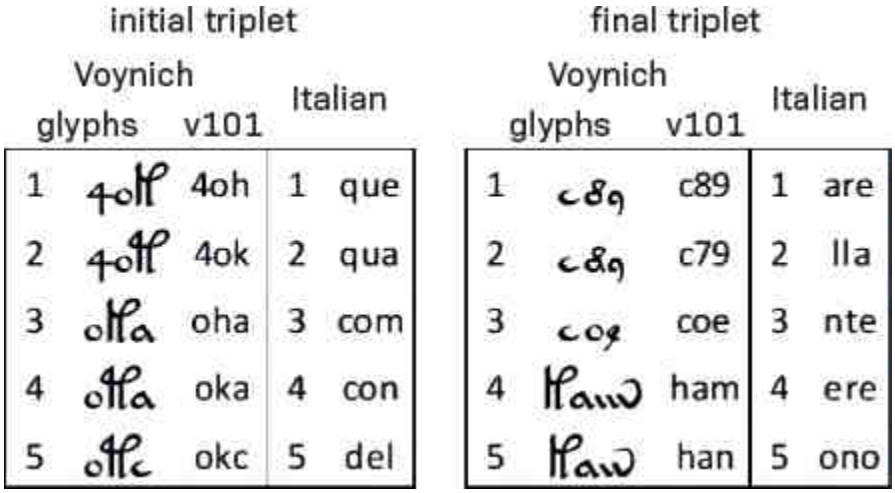
internal strings occurring over 100 times

final strings occurring over 100 times

Thus, these 100 glyph strings could be thought of as building blocks of the manuscript, accounting for 87.7 percent of all the glyphs. And, if the scribes were allowed to use single glyphs within words when glyph triplets and pairs were not available, then these 100 blocks could in principle generate essentially 100 percent of the manuscript.

We cannot say for sure that these strings were the fundamental building blocks of the Voynich manuscript, nor that the source documents were necessarily parsed in the way that we have imagined. However, it is instructive to compare a few of the top-ranked glyph triplets with the most-

common trigrams in, for example, medieval Italian. The five most common initial and final triplets, in v101 and in medieval Italian, line up as follows:[5]



initial triplet					final triplet				
	Voynich glyphs	v101		Italian		Voynich glyphs	v101		Italian
1		4oh	1	que	1		c89	1	are
2		4ok	2	qua	2		c79	2	lla
3		oha	3	com	3		coe	3	nte
4		oka	4	con	4		ham	4	ere
5		okc	5	del	5		han	5	ono

If this approach should ultimately yield a viable mapping, we can see firstly how it would solve the mystery of the inseparable glyph pair {4o} —namely, that this string could represent "qu" in the source language. Secondly, with a little imagination, we can make these blocks fit together. For example, the common glyph strings {4ohc89} and {4ohc79}, {4ohCC89} and {4ohCC79} could map in some permutation to the Italian "quello" or "quella" ("that one") and "quelli" or "quelle" ("those ones).

To put this hypothesis to a more rigorous test would require a working assumption on the source language; a set of frequency-based mappings between glyphs and letters, glyph pairs, and digraphs, and glyph triplets and trigraphs; and some trials on selected chunks of text. Such a test is eminently possible, and would be a productive area of research.

A "WORD"-TO-WORD-MAPPING

Since we do not have a general mechanism to parse the glyph strings, we could consider the possibility that the strings do not need to be parsed.

Perhaps we might find a mapping that is not between single or multiple glyphs and letters, but between glyph strings—what we have called "words"—and real words. Perhaps the scribes of the Voynich manuscript had on their tables a reference chart containing a column of most of the common words in the source language, and a corresponding column of glyph strings by which they were instructed to transcribe the source document, but in which the individual glyphs had no meaning.

We could call this a word-to-"word" mapping (in the sense in which the Voynich scribes would have perceived it); or a "word"-to-word mapping (in the sense that we are trying to reverse-engineer their work and return to the original documents).

There are least three objections to the hypothesis of a "word"-to-word mapping:

- We have no way of knowing whether the scribes preserved the lengths of words in the source documents. For example, they could have transcribed two-letter words to five-glyph strings, or vice versa. In response, we can take only the simplest hypothesis—that the lengths of words were preserved—and see where it takes us.
- Accepting for the moment that word lengths were preserved: the average length of glyph strings is very short, under four glyphs. As we have seen, the average length of words in most European languages is substantially longer. A possible response to this objection is that the source documents for the Voynich manuscript might have had unusually short words: as would be the case for a typical play by William Shakespeare, for example *Macbeth* (average 4.26 characters),[6] or even Dante's *La Divina Commedia* (average 4.22 characters).[7]
- The vocabulary of the Voynich manuscript is relatively small. We have estimated that Language A has at most 3,716 distinct "words," and Language B has at most 5,773; in both cases possibly less, depending on the extent to which the "rare glyphs" are distinct from each other. By comparison, Dante's *La Divina Commedia* has approximately 14,456 distinct words, and Bocaccio's *Il Decamerone* has approximately 20,072.[8] Again,

one could counter with the proposal that the source documents might have had small vocabularies: for example, *Macbeth* has 3,492 distinct words, and *A Midsummer Night's Dream* has only 2,931.[9]

In any case, our methodology for examining this hypothesis was as follows:

- We started with a working assumption that the source language for the Voynich scribes (and therefore our target language for any transliteration of the glyphs) was medieval Italian.
- We identified, as best we could, the most-common words of one, two, three, four and five letters in medieval Italian, drawing upon the OVI corpus as well as the full text of *La Divina Commedia* and *Il Decamerone*.
- For each such word, we counted the number of occurrences in the OVI corpus and ranked the words in order of frequency.
- Turning to the Voynich manuscript, treating Languages A and B separately and testing both the v101 and the v104 transliteration: we identified and counted the "words" of one, two, three, four, and five glyphs, and ranked the "words" in order of frequency.

The results, placing medieval Italian and the Voynich manuscript side by side, are summarized in Table 8.10.

With two descending sequences and relatively few points of data, it is not difficult to get good correlations. At each word length from one letter to five letters, there is a set of common words in medieval Italian, and there are sets of Voynich "words" in Languages A and B (generally not the same ones) whose frequencies correlate quite well with those of the Italian words. The correlations are generally preserved whether we use the v101 transliteration or our own, highly compressed, v104 alphabet. These results tend to support our assumption that, if the scribes used a word-to-"word" mapping, then they preserved the word lengths.

Two observations are in order.

The first is that in order to achieve a good correlation is between the single-letter words in medieval Italian, and the single-glyph "words" in the Voynich manuscript), we had to amalgamate the glyphs ⸗ [e] and ⸘ [Φ]. In all other circumstances, we had considered these glyphs to be distinct. We wondered therefore if the glyphs ⸗ {e} and ⸘ {9} were the same.

Secondly, there is an unusually low correlation—just 53 percent—between four-letter words in Italian and four-glyph strings in the v104 transliteration of Language A. This is a consequence of our reassigning the ubiquitous ⸸ ⸽ from {8am} in v101 to [8aîΩ] in v104, which converts it from a three-glyph to a four-glyph string. This in turn followed from our breaking down the glyph ⸽ {m} into two glyphs ⸽ [îΩ]. If we reverse that breakdown, the four-letter/four-glyph correlation jumps back to 96 percent.

Table 8.10. The most common words of one to five letters in medieval Italian, and the most common "words" of one to five glyphs in the Voynich manuscript

medieval Italian			Voynich Language A			Voynich Language B		
word	%	English	"word"	v104 key	%	"word"	v104 key	%
one-letter words				RSQ =>	94.4%		RSQ =>	94.7%
E	57.4%	and	⸦	s	33.3%	⸗	e	19.0%
A	21.2%	at, to	⸘	Φ	12.9%	⸘	Φ	17.1%
È	11.0%	is	⸮	Ψ	11.7%	⸮	Ψ	15.5%
I	4.5%	the (pl.)	⸆⸜	Δ	11.0%	⸦	s	13.9%
O	4.4%	either, or	⸸	8	7.6%	⸜	o	9.9%
two-letter words				RSQ =>	66.9%		RSQ =>	86.7%
DI	14.0%	of, than	⸜⸗	oe	10.5%	⸜⸗	oe	22.2%
LA	11.9%	the (f.sing.)	⸸⸘	8Φ	10.2%	⸳⸮	aΨ	16.0%
IN	7.8%	in	⸲⸘	1Φ	9.3%	⸜⸮	oΨ	12.3%
DE	5.0%	of	⸜⸮	oΨ	9.0%	⸳⸗	ae	11.9%
LO	5.0%	the (m.sing.)	⸰⸲⸜	1'o	8.6%	⸸⸘	8Φ	7.8%
three-letter words				RSQ =>	94.2%		RSQ =>	94.9%

CHE	22.6%	which, who	⟃⟃o9	1oe	9.6%	a ıı ᘐ	aîΩ	13.0%
PER	13.0%	for	⟃⟃oʔ	1oΨ	6.5%	⟃⟃c9	1cΦ	6.6%
NON	9.5%	not	a ıı ᘐ	aîΩ	4.9%	⟃ʔc9	1'cΦ	5.2%
DEL	5.7%	of the	⟃ʔo9	1'oe	4.1%	8aʔ	8aΨ	5.1%
CON	3.6%	with	8aʔ	8aΨ	3.6%	a ᘐ	aiΩ	3.9%
four-letter words				RSQ =>	61.8%		RSQ =>	93.5%
COME	7.5%	as/like, how	8a ıı ᘐ	8aîΩ	13.8%	⟃⟃89	1c8Φ	8.1%
LORO	6.2%	their, them	8a ᘐ	8aiΩ	2.7%	⟃ʔ89	1'c8Φ	6.7%
SONO	4.3%	are	⟃⟃o9	1coe	2.0%	8a ıı ᘐ	8aîΩ	4.9%
COSE	3.9%	things	⟃⟃oΨ	1coΨ	1.4%	⟙olf⟃⟃9	ΔhĉΦ	3.2%
COSA	3.9%	thing	⟙olf⟃⟃9	Δk1Φ	1.4%	⟙olfa9	Δhae	2.6%
five-letter words				RSQ =>	59.4%		RSQ =>	80.0%
DELLA	5.7%	of the	olfa ıı ᘐ	ohaîΩ	1.7%	⟙olfa ᘐ	ΔhaiΩ	4.0%
QUALE	5.7%	which	olf⟃o9	ohcoe	1.7%	⟙olf⟃89	Δhc8Φ	4.0%
QUALI	3.4%	which (pl)	⟙olf⟃o9	Δhcoe	1.6%	⟙olfa ıı ᘐ	ΔhaîΩ	3.3%
DETTO	2.4%	said	⟃olfa ıı ᘐ	okaîΩ	1.6%	⟙olf⟃⟃89	ΔhĉΦ	3.0%
DISSE	2.2%	he said	⟃⟃o89	1co8Φ	1.5%	olfa ıı ᘐ	ohaîΩ	2.3%

Source: author's analysis based on OVI corpus of medieval Italian, and v104 transliteration of the Voynich manuscript. For each word length, RSQ denotes correlations between the counts of all Italian words and the counts of all Voynich glyph strings of that length (it is not necessarily implied that the glyph strings correspond to the respective words).

Indirectly, therefore, we may have found evidence d that ıı ᘐ [îΩ] is, after all, a single glyph and not a string. That will serve us when we return to the concept of a glyph-to-letter mapping.

With these reservations, it would be tempting at this point to say that we might have found a mapping that applies only at the level of words, and not at the level of glyphs or letters.

It would be an extraordinary coincidence if the frequencies of the short Voynich "words" were a close match for the frequencies of the short words in Italian, or indeed in any language. However, as an experiment, we applied these mappings to the whole of the text in Languages A and B,

using our v104 alphabet. This test did indeed produce a few lines in which every, or nearly every, Voynich "word" was replaced by an Italian word of the same length. For example, we had an almost complete mapping of line 25 on page f075r, which Currier had identified as being in Language B. This line was in two parts, with an intervening illustration, as shown below:

In our v104 transliteration, this line read as follows:

saiΩ 2ĥΦ Δhĉ8Φ 2Φ 8Φ -- Δhĉ8Φ e1c8Φ ΨaiΠ

where -- denotes the break in the text. In our mappings this line became, in medieval Italian:

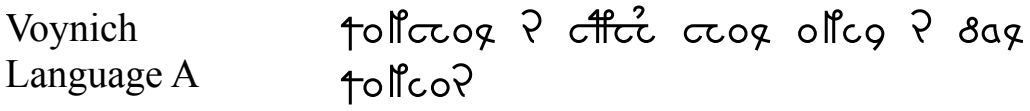
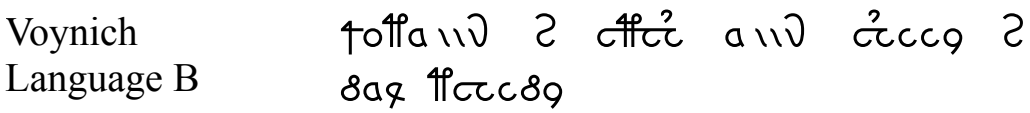
ANNO VOI DETTO LI LO -- DETTO TERRA [ΨaiΠ]

where the final word ʔɑɣ [ΨaiΠ} was too infrequent to permit a mapping.

The above sequences of medieval Italian words do not make much sense in English. However, we could imagine that, with a little swapping of words with similar frequencies, we might arrive at a sensible sentence in one language or the other.

Taking the mapping in the reverse direction: if we imagine for a moment that the source document was *La Divina Commedia*, we can find in this epic poem many lines which use several common short words in medieval Italian. If, purely hypothetically, we treat the frequency rankings in Table 8.10 as a real mapping, we can create Voynich equivalents of Dante's immortal lines. Here is an example from Canto I (with the infrequent Italian word "amara" left unmapped):

Italian Tant' è amara che poco è più morte

English equivalent	So bitter it is, that death is little more
Voynich Language A	
Voynich Language B	

These invented “sentences” do not appear anywhere in the Voynich manuscript. It would be extraordinary if they did. But they look as if they could.

The downside of a “word”-to-word mapping is that frequency analysis will break down when the “words” become rare. When we were thinking about a glyph-to-letter mapping, we had to worry only about the “rare” glyphs which, in total, occur only 341 times in the whole manuscript. But when we are dealing with “words,” we have to consider that there are 6,714 “words”—16 percent of all the “words” in the manuscript—that occur just once. If their constituent glyphs give us no clue to their meaning, the best we could hope for is that we could guess those “words” from their context.

Nevertheless, the above exercises show that a “word”-to-word mapping might be worth study. A proper test of the hypothesis would require a much more complete mapping between words in selected medieval languages, and glyph strings in the Voynich manuscript. And if the first selected languages did not work, then the same exercise would have to be repeated with other medieval languages. Like quite a few tasks that we have looked at, that one is beyond the scope of this book.

GLYPHS TO DIGRAPHS

As a matter of due diligence, while we are examining strings of glyphs and letters, we have at least to consider the possibility that single glyphs could map to pairs of letters (digraphs) in natural languages.

If the scribes of the Voynich manuscript had used this fiendish mechanism, it would help explain at least the following phenomena:

- that the “words” in the Voynich manuscript are extremely short (averaging about four glyphs)
- that there are several glyphs that occur predominantly at the ends or beginnings of “words”
- that the Voynich “alphabet” seems unusually large (at least if we accept all the variations that are defined by the v101 keyboard assignments)
- that there are dozens of “rare glyphs” that occur only once or a few times in the whole manuscript

If one glyph maps to two letters in a natural language, then the language in question will have an average word length of about eight letters. Referring to Table 6.2 in chapter 6, we can see that among modern European languages, the closest matches would be Ukrainian, Latvian, and Belarusian; but Latin is also a good match.

In natural languages, many letter pairs are prefixes or suffixes. For example, in modern English, the thirty most common letter pairs include “in,” “re,” and “en,” which are often prefixes, and “er,” “ed,” “es,” and “ng,” which are often suffixes or parts of suffixes. So, if single glyphs map to letter pairs, and if spaces map to spaces, some of the glyphs will show up frequently at the beginnings or ends of “words.”

A mapping of this nature needs a large number of different glyphs, in order to address the wide range of possible letter pairs. Among the European languages, there are typically between thirty and forty letter pairs that have a frequency of at least 0.75 percent (that is, they are at least as common as the letter “z” in English). For example, in modern English the most common letter pair is “th” with a frequency of 4.0 percent; the thirtieth most common pair is “ea” with a frequency of 0.8 percent.

However, in any natural language, there will be many more letter pairs that occur less frequently but will still have to be accommodated in a glyph-to-digraph mapping. We could imagine that what Glen Claston called the “rare glyphs” were created for exactly this kind of eventuality.

We can foresee at least the following objections to this hypothesis:

- The frequency distribution of letter pairs is very flat. As we saw above, in English the letter pair “th” is only five times more frequent than the thirtieth pair “ea.” The distributions have a similar “flatness” in most European languages. in the Voynich manuscript, the most common glyph {o} is over 100 times more frequent than the thirtieth, {f}. So, if the manuscript was mapped from digraphs in European languages, it has far too many instances of {o}, and probably also of {9} and other “super-glyphs.”
- The scribes of the Voynich would have needed a mechanism to transcribe words with an odd number of letters. For example, the common English word “there” starts with two common digraphs, “th” and “er,” for which we can imagine the scribes having corresponding glyphs at the ready. But then there is an orphan letter “e,” which will demand its own glyph.
- Even single-letter words, such as “a” and “I” in English, or “a,” “à,” “e,” “è,” “i” and “o” in Italian, would need glyphs of their own. Most European languages have no more than six distinct single-letter words, including accented variants. We might therefore have to postulate that the scribes reserved up to six glyphs for single-letter words; in that case, those glyphs would have to be {s}, {8} and its variants, {9}, {y} and its variants, {e} and {o}, in some order. They could then reuse these “reserve” glyphs whenever the same letters were “orphans” at the end of words. But they would still need glyphs for other “orphan” letters. For example, the Italian word “per” would need a glyph for the digraph “pe” and another for the letter “r,” which does not occur as a word and therefore could not use one of the “reserve” glyphs.
- There would have to be a separate mechanism for words of two letters in the source document. Clearly, single glyphs cannot map to two-letter words since there are not enough glyphs. The Voynich manuscript uses only forty-four glyphs as single-letter “words,” whereas medieval Italian has at least seventy distinct two-letter words[10] and old French has at least 155.[11]

In any event, if we are to hypothesize a glyph-to-digraph mapping, we have to put it to the test. The logical test is still the "Gold Bug"; if a letter pair is very frequent, it should map to a very frequent glyph, and vice versa. Only a mapping of real text will test the hypothesis.

We therefore selected the ten most frequent "words" in Language A, and the ten most frequent in Language B.

In Tables 8.11A and 8.11B are our mappings of these "words" into selected European languages, on a simple matching of glyph frequencies with digraph frequencies.

Most of the mappings do not yield real words in the candidate languages, but some of them do.

To take this analysis further would require a deeper look at the concepts of "reserve" glyphs and "orphan" letters and would benefit from some computing power. Testing the possible combinations would stretch the capacities of Microsoft Word and Excel, which are not designed for this kind of work. We feel that the hypothesis of glyph-to-digraph mapping will repay further study. Interested readers are invited to step forward.

Table 8.11A. Voynich manuscript, v101 transliteration, Language A, top ten words: mappings from single glyphs to digraphs in selected languages

		German, modern	Croatian, modern	Spanish, modern	Czech, modern	Latin, classical
R squared=>		94.9%	94.7%	93.5%	91.7%	86.1%
"word"	v101					
ꞩa ıı𝔍	8am	IEDEND	STKOTA	LAESAR	JEPORO	REQUEN
ƈƈoϥ	1oe	CHENEI	RAJENI	ERENUE	CHSTNA	INERIS
ƨ	s	ES	EN	EL	SE	TU
ƈƈoʔ	1oy	CHENGE	RAJEPO	ERENNT	CHSTBY	INERUS
oϥ	oe	ENEI	JENI	ENUE	STNA	ERIS
a ıı𝔍	am	DEND	KOTA	ESAR	PORO	QUEN
ƈƈ9	19	CHER	RANA	ERDE	CHEN	INUM
oʔ	oy	ENGE	JEPO	ENNT	STBY	ERUS
ꞩ9	89	IEER	STNA	LADE	JEEN	REUM

⁊oꬷ	2oe	BEENEI	VAJENI	TAENUE	LASTNA	TEERIS

Source: author's analysis; digraph frequencies in selected languages from Stefan Trost Media

Note: R squared is the correlation between frequencies of the forty most common glyph pairs in the v101 transliteration, Language B, and the frequencies of the most common digraphs in the candidate languages.

identifiable word in the candidate language

WHERE NEXT?

At this point we have to set aside our analysis of glyph frequencies. Perhaps we must look for a more intuitive approach. As codebreakers say, we need a crib. As we noted in an earlier chapter, we may find such a crib among the herbal pages or the zodiac pages, where we might reasonably hope to find the unmistakable name of a plant or the imperishable name of a constellation. We shall narrate this search in the next two chapters.

Table 8.11B. Voynich manuscript, v101 transliteration, Language B, top ten words: mappings from single glyphs to digraphs in selected languages

		French, modern	Spanish, modern	Catalan, modern	Galician, modern	Latin, classical
R squared=>		97.3%	93.9%	93.4%	91.9%	89.2%
"word"	v101					
oꬷ	oe	ESOU	ERRA	ERAR	DETE	INNT
ɑıı∀	am	DEAN	ESQU	DERA	RACO	QUAT
⊂⊂ɕ9	1c89	NTENRELE	OSDEUEEN	REENELES	NTERESEN	TIUMISER
ɑʔ	ay	DEER	ESTE	DETA	RADO	QUIT
oʔ	oy	ESER	ERTE	ERTA	DEDO	INIT
ɕɑıı∀	8am	REDEAN	UEESQU	ELDERA	ESRACO	ISQUAT
⊂⊂9	1c9	NTENLE	OSDEEN	REENES	NTEREN	TIUMER
ɑꬷ	ae	DEOU	ESRA	DEAR	RATE	QUNT

ϯοlℱαιℐ	4ohan	ONAIDEME	NTLAESCO	QULADEPE	OSARRAQU	USREQUES
ϱ	e	OU	RA	AR	TE	NT

Source: author's analysis; digraph frequencies in selected languages from Stefan Trost Media

Note: R squared is the correlation between frequencies of the forty most common glyph pairs in the v101 transliteration, Language B, and the frequencies of the most common digraphs in the candidate languages.

identifiable word in the candidate language

CHAPTER 9

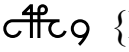
CANNABIS

If we are to find a crib for the Voynich glyphs, surely our first port of call will be among the herbal pages. Every one of these 129 pages contains at least one illustration of a plant, nearly always occupying the full height of the page (but not necessarily the width of the page). Every illustration is accompanied or interspersed by lines or strings of glyphs that look like text. Nearly every illustration is colored, and the colors of the flowers in particular ought to give us clues to the identity of the plant. Following Occam's razor, our simplest assumption is that somewhere among these glyphs is a representation of the name of the plant, in some language that we have to identify. In chapters 5 and 8 we explored the "Gold Bug" approach, which, if it has not broken the bars of the manuscript, has at least given us a means of prioritizing the languages that we shall test.

FLAGS

But before we dive into the search for a crib, let us consider what we should be searching for.

In Chapter 7, Table 7.6, we saw that each of the seven thematic sections has some glyph strings or "words" that we could call flags for the section. By flags, we mean that such a "word" is among the twenty most common "words" in the section, but rarely occurs in any of the other sections. These "words," if they are really words, could be thought of as representing concepts that are specific to the section.

In the case of the herbal section, we saw that the flag strings included  {K9}, {2o), {8an} and {1o). Other than these, the most-common "words" in the herbal sections were "words" that occurred, in all the other sections. Here we run into a conundrum. If the text of the herbal section is really related to the bold and fantastical illustrations that populate the section, we should expect some of the flag strings to represent concepts like "plant," "flower," "leaf," "stem," or "root." But in what natural language are these concepts represented by words of two or three letters?

These short strings, then, do not seem like good candidates for botanical keywords. We need to search among the longer strings. Furthermore, we should look for strings that occur often in the herbal section but rarely in the sections that we expect to be unrelated to plants (for example, the biological/balneological section). For good measure, we should also search for the same strings in the pharmaceutical section, where there are many drawings of what look like the roots and leaves of plants. If these strings occur there, even if less frequently, we might hope that they represent botanical concepts.

In Table 9.1 is a tabulation of the ten most common "words" of four, five and six glyphs in the herbal section, and the extent to which they also occur in the pharmaceutical, biological, and cosmic sections.

Here we can see that there at least ten "words" of between four and six glyphs which are common in the herbal pages but are rare or nonexistent in the biological and cosmic pages. Some of them have also a few occurrences in the pharmaceutical section, which is as we should expect if they represent botanical terms.

The only other observation that we can make is that four of these "words" begin with ꟈoꟊ [Δh] or ꟈoꟗ [Δk]. We could reflect on whether these strings represent prefixes in the normal linguistic sense; or are some kind of abbreviation, analogous to the term "sp." ("species") in modern biological terminology.

Table 9.1. Voynich manuscript: ten possible keywords for the herbal section

"word"	v104 key	"word" count, by section			
		herbal	pharmaceutical	balneological/biological	cosmic
ꟈoꟗcc9	Δk1Φ	48	2	3	0
cco89	1o8Φ	44	8	0	5
ꟈolꟗcc9	Δh1Φ	44	1	5	2
ccco2	1coΨ	32	20	10	12
oꟗcc9	ok1Φ	32	0	2	6
ꟈolꟊo4	Δhoe	32	22	25	5
oꟗo4	okoe	31	7	13	15
cco8a\\J	1o8aîΩ	24	5	0	0
oꟗcco4	ok1oe	22	2	0	1
ꟈoꟗo2	ΔkoΨ	16	0	1	1

Source: author's analysis. "Cosmic" includes the astronomy, cosmology and astrology/zodiac sections.

It is also possible that these initial strings are simply markers with a grammatical function or some other function unrelated to the meaning of the text.

As it stands, the above analysis tells us only that there are glyph strings that are unique or nearly unique to the herbal section. With this, we have at least some justification to search for the names of specific plants, even if we expect such names to occur only once.

THE VOYNICH PLANTS

Given the wealth of material that is available to the researcher, it is quite surprising that relatively few people have worked seriously on the identification of the plants. Following René Zandbergen's summary on his excellent website, we are aware of the following:

- Hugh O'Neill (1894–1969), a botanist at the Catholic University of America, attempted identifications of some of the Voynich plants. His original work is hard to find, but his colleague Theodore C. Petersen quoted at least thirteen of these identifications. In 1944, O'Neill published his views that two of the illustrations represented respectively the sunflower (*Helianthus Annuus L.*) and the pepper plant (*Capsicum* sp.), both genera native to the Americas. He concluded that the illustrations dated from later than the return of Cristoforus Columbus from the Americas to Europe in 1493.[1] If that were so, the manuscript would have been at least partly written several decades after its parchment was produced (which, to this author's mind, should not be thought extraordinary).
- Herman Theodor Holm (1854–1932), whose main career was as a plant pathologist for the US Department of Agriculture, was cited by Elizabeth Friedman in 1962 as having identified sixteen plants as European species. Ms. Friedman further quoted Holm as follows: "The remainder are composite: i.e., the root system belongs to one plant, the stem system to another, the leaves, and flowers to still others. A few show imaginary root or flower structures."[2] This seems to have been the first scientific assessment to the effect that the Voynich plants, or at least some of them, were fantasies.
- Theodore C. Petersen (1883–1966), a professor of ancient languages and history at the Catholic University of America (and evidently a colleague of both Hugh O'Neill and Herman Holm), worked intensively between 1931 and 1944 on an attempted decipherment of the Voynich manuscript. In the course of this work, he compiled tentative identifications of 102 of the plant illustrations. Mary D'Imperio later recalled that "The list of plant identifications compiled by Petersen in his

hand transcript includes identifications [that] he attributes to Mr. [Wilfrid] and Mrs. [Ethel Lilian] Voynich, [Hugh] O'Neill. and [Herman Theodor] Holm."[3]

- Ethel Lilian Voynich, according to René Zandbergen, "appears to have used Th[eodore] Petersen's lists and other sources to create her own list of tentative identifications." These lists are held at the Grolier Club Library in New York and at the Beinecke Library in New Haven.[4]
- The most comprehensive identification of the Voynich plants is surely that of Dr. Edith Sherwood, who on her website[5] presented her names of 125 of the plants, in most cases accompanied by a photograph of what she considered a modern equivalent.

As far as we can determine, none of these researchers made any use of the accompanying glyphs as an aid in their identifications.

This author is not a botanist and has no means of evaluating these or any other identifications of the Voynich plants. That task, like others that we would have liked to attempt, is beyond the scope of this book. What we can do, however, is to examine to what extent the various identifications agree or disagree (excluding those which explicitly cite another author). Among the sources listed above, we found only ten plants on which more than one author was in agreement. These are listed in Table 9.2. Nine of the corresponding pages were in Currier's Language A and one was in Language B (though his identification of these languages was first made known in 1976, decades after these identifications were made).

This table, which documents some measure of agreement between researchers with different backgrounds, and encompasses a long period of research, seems to us a reasonable starting point to look for cribs in the glyphs that accompany the plant illustrations.

If for no better reason than that the leaf is familiar to everyone who remembers the 1960s, we propose to start with page f016r. With regard to the illustration on this page, Theodore Petersen and Edith Sherwood agreed that it depicted a species of the *Cannabis* genus, or Indian hemp.

Dr. Sherwood is forthright in her identification of the plant:

Folio 16r, Cannabis, is indigenous to Central Asia around the Caspian Sea. It was introduced into Western Asia and Africa, and subsequently, Europe in 2,000–1,000 BC. Cultivation in Europe became widespread in about 500 AD. Cannabis is one of the world[']s oldest crops and provides a number of useful products: fiber from the stems, therapeutic agents and narcotics from the flowers and leaves, and oil from the seeds.[6]

Table 9.2. Voynich plant illustrations with common identfications by more than one researcher

		author of plant identification			
page	Currier language	Theodore Petersen Latin name	Ethel Lilian Voynich Latin/English name	Edith Sherwood English name	Edith Sherwood Latin name
f001v	A	*Solanum, Solatrium, Belladonna*	*Atropa Belladonna* +	-	cites Petersen
f002r	A	-	*Cyanus?*	Diffuse Knapweed	*Centaurea diffusa*
f004r	A	*Hypericum, Centarium Erythaea* (O'Neill)	*Hypericum?*	Alpine Saxifrage	*Saxifraga cespitosa*
f009v	A	*Viola tricoloris, Herba trinitatis* (O'Neill)	Viola; pansy	Viola	*Violaceae*
f015v	A	*Herba Paris*	*Lilium?*	Herb Paris	*Paris quadrifolia*
f016r	A	*Canabis* (hemp)	*Aconitum?*	*Hemp?*	*Cannabis*
f032v	A	*Campanula, Rapunculus, Convolvulus erectus?*	*Malva?*	Harebell	*Campanula rotundifolia*
f053v	A	*Hieracium* (hawk weed)	-	Hawkweed	*Hieracium aurantiacum*
f087v-left	A	*Primula (dianthus);*	-	Primrose	*Primula vulgaris*

f094r	B	*Botriculum, (Lunaria?* Mondfarn?)	-	Moonwort or Moonfern	*Botrychium lunaria*

Source: Holm, Petersen and Voynich from Zandbergen, Rene (2016), Pre-1960's tentative herb identifications; Sherwood from Sherwood, Edith, PhD and Erica Sherwood, *The Voynich Botanical Plants* (2015).

There has been some disagreement among scholars as to whether the *Cannabis* genus has a single species or several. The US National Plant Germplasm System[7] identifies three main species: *Cannabis indica* Lamarck, *Cannabis ruderalis* Janischevsky, and *Cannabis sativa* Linnaeus.

Part of the illustration, and the whole of the all-important text, is reproduced in Figure 9.1. We observe that the leaf has seven leaflets, which is consistent with *Cannabis indica, Cannabis sativa*, or *Cannabis ruderalis*; and that the leaflets are quite wide, which suggests *Cannabis indica*. However, we can reserve judgment on whether the spiky leaves and the tiny flowers resemble the plant that powered the "Summer of Love." Our quest is to find a word for this plant in a European language among the glyphs that accompany the illustration.

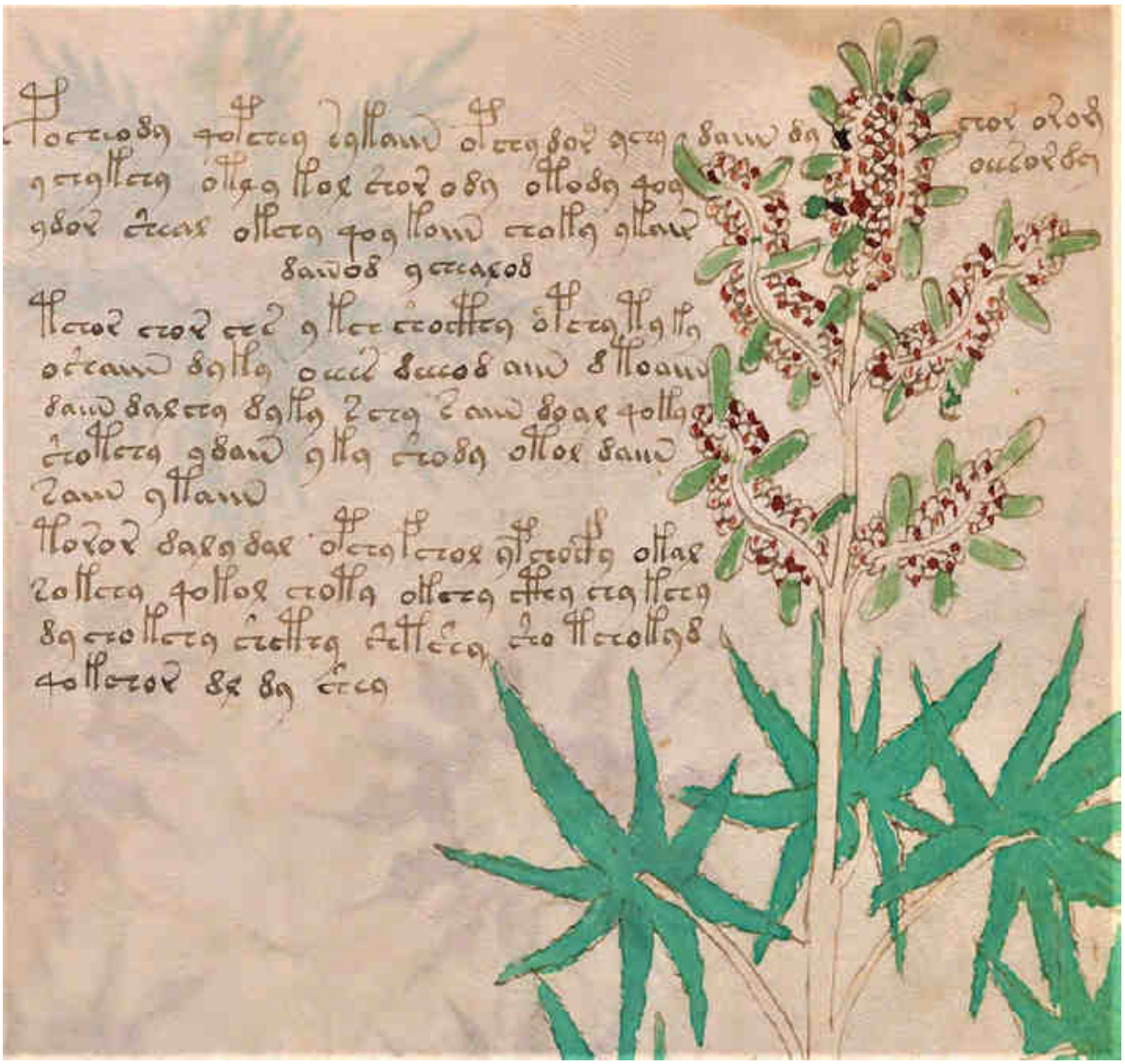


Figure 9.1. Voynich manuscript, detail from page f016r, including the full text and part of the illustration. *Beinecke Rare Book and Manuscript Library*

The cannabis plant is known in vernacular English as hemp. Thanks to the Multilingual Multiscript Plant Name Database (MMPND)[8] at the University of Melbourne, and to Google Translate, we have the vernacular names in at least the following other European languages:

Czech	konopí
Danish	hamp, hampslægten
Dutch	hennep

Finnish	hamppu
French	chanvre
Galician	cánabo
German	hanf
Italian	canapa
Portuguese	cânhamo, maconha
Spanish	cáñamo
Swedish	hampa, hampor
Turkish	kenevir
Ukrainian	конопель

So, if the illustration on f016r is really the cannabis plant, then somewhere among the glyphs we should hope to find a string that might be persuaded to conjure one or more of these European words.

Let us start with Galician, Portuguese, and Spanish, which have similar structures. We are then looking for a string of six or seven glyphs in which the second and fourth glyphs are the same (or almost the same), and the last glyph is a vowel which should be different from the second and fourth.

On page f016r, there is only one candidate for such a glyph string. It is the first “word”:

{go1co89}

If this “word” could be transliterated as “cánabo,” “cânhamo,” or “cáñamo,” we would have the following crib:

{g} = “c”

{o} = “a” or “â” or “á”

{1c} = “n” or “ñ” or “nh,” in which case the mini-glyph that looks like a {c}, is really just an accent

{8} = “m” or “b”

o {9} = “o.”

If this crib works, it will give us a transliteration of four of the eight most common glyphs in the Voynich manuscript. We should then, if we resurrect the “Gold Bug” approach, have a reasonable chance of transliterating other glyphs which do not show up in the crib, on the basis of their frequencies in the manuscript as a whole.

For example, if we think that the language of f016r is Spanish (and that therefore the first word is “cáñamo”), we could try transliterating some other common glyphs on the basis of letter frequencies in Spanish.

We expanded the crib to a provisional transliteration of the ten most frequent “words” in Language A, as shown on Table 9.3.

Most of these mappings did not result in real words on Spanish, but a few made sense.

Our second shot was the modern Czech language, which we took as an approximation to medieval Bohemian. The results were the same.

If we now turn to the Italian “canapa,” we need a string of six or seven glyphs in which the second, fourth, and last glyphs are the same. There is no such string on page f016r.

Likewise, if we are searching for the German “hanf,” the French “chanvre,” or their equivalents in other languages, there is no obvious candidate on page f016r for the crib that would “break the bars” of the transliteration.

If a crib works for one word, it has to work for a reasonable number of words; otherwise, it is not a crib. Notwithstanding Theodore Petersen’s and Edith Sherwood’s accord on the identification of the plant, we have not found the name that we hoped to find, at least not in any of the languages that we have tried.

Table 9.3. Voynich manuscript, v104 transliteration, Language A: a “crib” mapping of the top ten “words”

“word”	v104 key	Spanish	Czech
“crib			
	go1co8Φ	CÁÑAMO	KONOPI

the ten most frequent "words"

ϑα ιιυ	8aîΩ	MSDU	PADV
cco4	1oe	NAR	NOT
ƨ	s	P	M
ccoʔ	1oΨ	NAI	NOL
o4	oe	AR	OT
α ιιυ	aîΩ	SDU	ADV
cc9	1Φ	NO	NI
oʔ	oΨ	AI	OL
ϑ9	8Φ	MO	PI
c̓co4	1'oe	N?AR	N?OT

Sources: glyph frequencies by author; letter frequencies by courtesy of Stefan Trost Media; reference corpora were www.corpusdelespanol.org (Spanish) and *Dalimilova kronika* (Czech).

common words in respective corpora

common strings in respective corpora

rare strings in respective corpora

CARCHUELA

Before we leave the search for "cannabis," we may be permitted to narrate our research a little out of sequence.

In Chapter 13 we shall explore the hypothesis that one of the underlying languages of the Voynich manuscript is medieval Galician-Portuguese. On this hypothesis, and according to our most plausible mapping between glyphs and letters, the first glyph on page f016r could represent the word "*ot*osa," with great uncertainty about the first and fourth glyphs.

We found a reference to a plant known as "arbea lotuosa" which was reported to grow in certain olive plantations in Andalusia.[9] The label on a bottle of "Picual" olive oil depicts this flower; it is labeled with the Latin

name *Convolus arvensis* and the vernacular name "carchuela." In English-speaking countries, the plant with the Latin name *Convolvulus arvensis* is known as field bindweed. Typically, it has small flowers with alternating red and white stripes; is this what the artist intended with the red dots on page f016r?

Could then "arbea lotuosa," "carchuela," or *Convolvulus arvensis* be the plant depicted on page f016r? The illustrations in Figure 9.2 present a visual comparison.

MOON FERN

A reasonable next step would be, for the moment, to leave Language A aside, and to search for a crib in the one page in Language B where at least two researchers are known to have agreed on the plant identification. That page is f094r. Theodore Petersen thought that the plant on this page might be "Mondfarn," which approximates to the German for "moon fern," and he recorded the Latin name as *Botriculum lunaria* (with a question mark). Decades later, Edith Sherwood would come up with the same identification: Moonwort or Moonfern (*Botrychium lunaria*). Part of the illustration on page f094r is reproduced in Figure 9.3.

Figure 9.2. (*Left to right*) Voynich manuscript, page f016r, an extract from the plant illustration, *Beinecke Rare Book and Manuscript Library*; *arbea lotuosa* or carchuela plant (*Convolus arvensis*), Monteblanco Luis Pino SL; field bindweed (*Convolvulus arvensis*), *plantsan.com*

Now we need to search for the names of this plant in as many European languages as we can find. Here the Multilingual Multiscripted Plant Name

Database is no help to us, as it does not include the genus *Botrychium*. From the Global Biodiversity Information Facility (GBIF),[10] and from Wikimedia Commons, we can glean the following vernacular names:

Czech	vratička měsíční
Danish	almindelig månerude
Dutch	gelobde maanvaren
Estonian	kuu-võtmehein
Finnish	ketonoidanlukko
French	botryche lunaire, herbe à la lune, monnaie du pape
German	mondraute, echte mondraute, mond-rautenfarn
Hungarian	kis holdruta
Icelandic	tungljurt
Irish	lus na míosa
Swedish	Låsbräken
Ukrainian	гронянка півмісяцева
Welsh	lloerlys

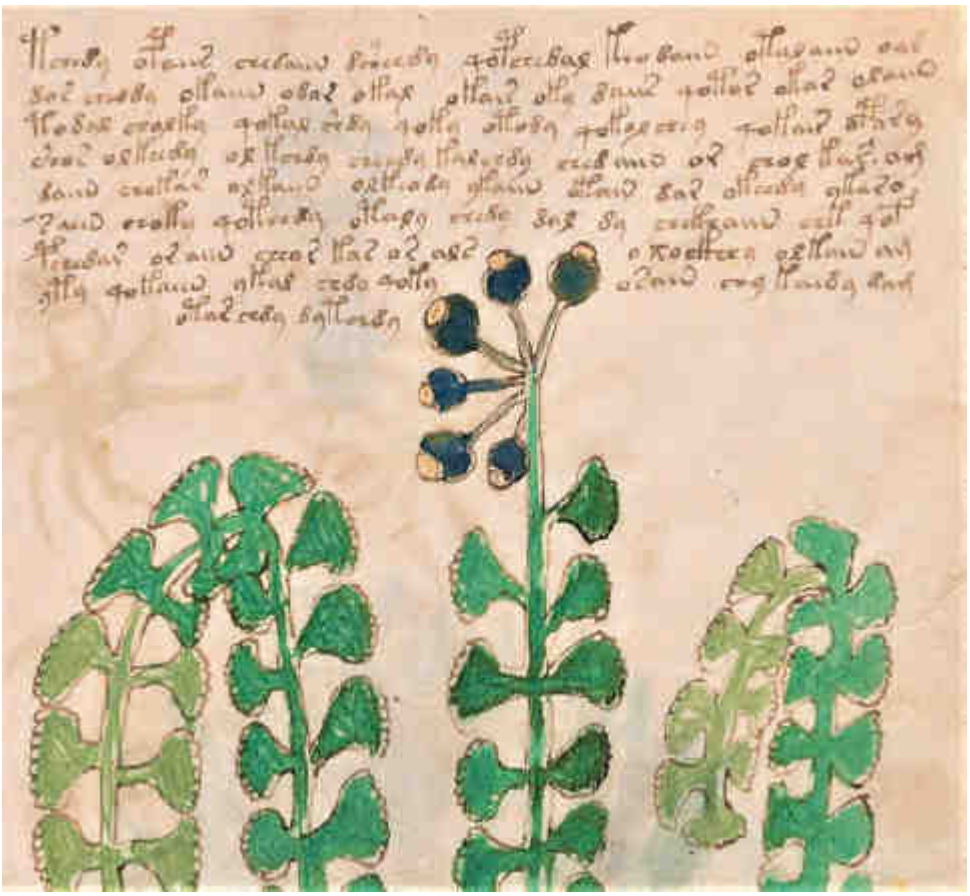


Figure 9.3. Voynich manuscript, detail from page f094r, including the full text and part of the illustration. *Beinecke Rare Book and Manuscript Library*

And there are more, in Arabic, Armenian, Bulgarian, Chinese, Lithuanian, Polish, Russian, and other languages, which (for the moment) seem to us implausible candidates for the basis of the Voynich text.

If any of these vernacular names is hiding among the glyphs on page f094r, it must surely be one of the shorter ones. We are tempted therefore to look for the Hungarian “kis holdruta,” the Welsh “lloerlys” or the Irish “lus na miosa.”

In Welsh, the digram “ll” is considered a single letter. It is not a doubled “l”; it is pronounced like “khl,” with the “kh” resembling the “ch” in the Scottish word “loch.” So if one of the underlying languages of the Voynich manuscript is Welsh, the “ll” will surely have its own glyph. Since “ll” is a relatively common digram in Welsh, the corresponding glyph should also be a relatively common one, If so, the Welsh “lloerlys” should be represented by a string of seven glyphs, with the first glyph a common one, and no

single glyph recurring. We should add that the first glyph should not be a vowel. This probably excludes the "super-glyphs" ᴄ [c], ᴏ [o] and ᴀ [a], which are so frequent that they surely must be vowels, and the glyph ᴓ [Φ], which we suspect to be a marker glyph with no meaning.

To this author's eye, there is no such string on page f094r.

In Irish, if we are hoping for one-to-one mapping of glyphs to letters, our search for "lus na miosa" should direct us to a sequence consisting of a three-glyph string, followed by a two-glyph string, and then a five-glyph string, all three strings beginning with consonants. The last glyph of the first string, and the fourth glyph of the third string, should be the same. Again, on page f094r there is no such sequence.

In Hungarian, the word "kis" means "small," so we should probably be looking for the word "holdruta," in which "hold" means "moon." In this case we are searching for a string of eight glyphs, beginning with a consonant and with three vowels, none of them repeated. We did not identify such a string on page f094r.

In short, the switch to Language B has not advanced our quest. Despite Theodore Petersen's and Edith Sherwood's agreement on the plant that is illustrated on page f094r, there is nothing among the glyphs on that page which gives us any confidence that we see the name of moonwort or moon fern, in any of the languages that we had proposed as candidates.

POSTSCRIPT: LUPINS

Before we leave the plant pages, this author feels impelled to recount a line of enquiry that was based on nothing more than instinct, or if we prefer, the triumph of hope over experience. It arose from a correspondence with another researcher of the Voynich manuscript. The subject of the correspondence was the plant on page f006v, wherein the text is in Currier's Language A. This plant has pointed leaves, each with seven leaflets, and flowers that look like spiky balls, at the end of long stems. In the illustration, the flowers have been colored green and then crudely daubed with circles of blue paint. It is not clear whether the artist intended the

flowers to be blue, or whether the blue circles were a hurried attempt to indicate that the spikes should be blue.

For what it is worth, Hugh O'Neill identified this plant as *Bidenus hypnis* (possibly he was referring to the genus *Bidens*, which has barbed fruits that stick to clothing); Theodore Peterson made tentative propositions of black gum tree (possibly referring to *Nyssa sylvatica*), or an *Arctium* [burdock] species, or the castor oil plant (*Ricinus communis*). Ethel Voynich accepted *Ricinus communis*. However, Edith Sherwood identified the plant as sea holly (*Eryngium maritimum*).

Our correspondence began as follows (VR denoting the Voynich researcher at the other end):

> VR: "The best exercise would be to try translating one of the plant pages…. This is a double test, as you'll need to identify the plant first and then translate the text to see if it describes the plant.[11]

This author replied with rash confidence as follows:

> RE: The page that you sent me is folio f6v. The plant is ricinus comunis.
>
> VR: … the plant is not R. communis. It has blue flowers, and the leaf lobes are fully divided. Clue: It is a Mediterranean food plant.[12]

Figure 9.4 illustrates some of the plants that this author considered in the quest for identification of the plant on page f6v. None of these earned a cigar. Our correspondence resumed as follows.

> VR: The species is *Lupinus micranthus* (Small-flowered Lupin).
>
> VR: The first line [of the glyphs] describes the important part of the plant: i.e. the edible crop.

Ricinus communis (castor bean plant)

Eryngium maritimum (sea holly)

Echinops ritro (Veitch's Blue)

Lupinus micranthus (small-flowered lupin)

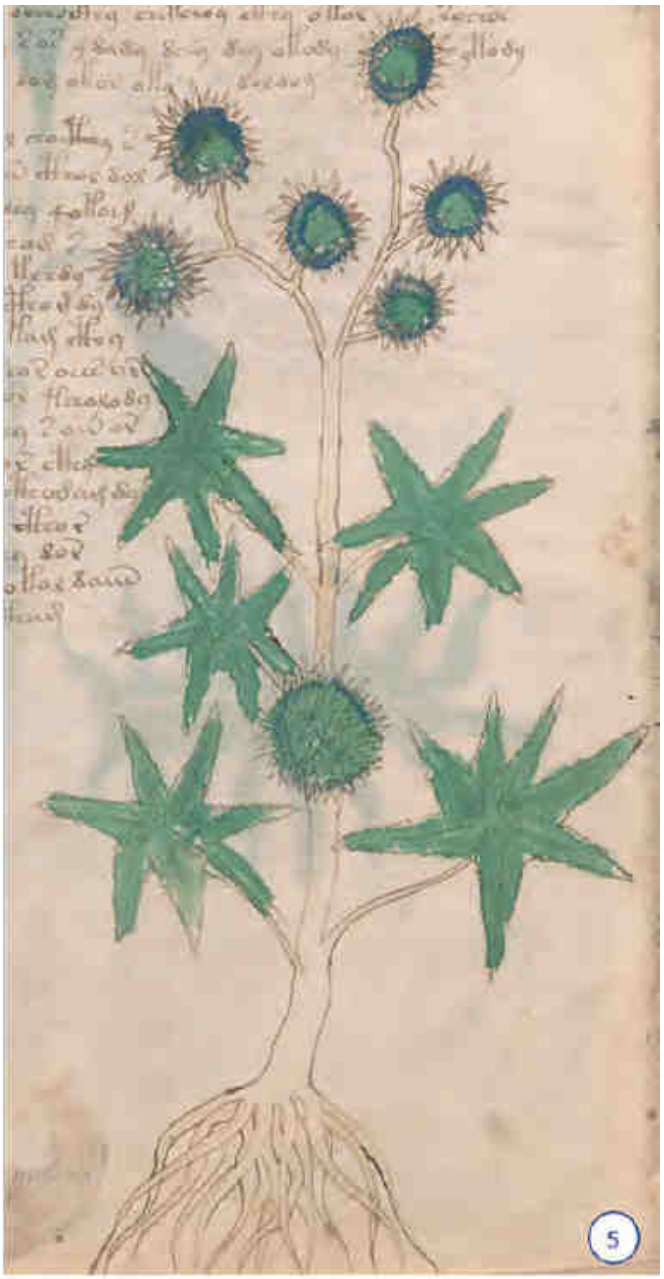


Figure 9.4. Voynich Manuscript: plant illustration on folio f006v, and possible modern counterparts. *(1) J. Young, (2) Alan Fryer, (3) Ballyrobert Gardens, (4) Abdelmonaim Homrani Bakali, (5) Beinecke Rare Book and Manuscript Library*

> VR: [the first three words, in Galician-Portuguese, are] "seedpod of noise," which is exactly what lupins have, as their seeds rattle in the pod.
>
> RE: Using your key, I attempted a complete transliteration of page f6v into Roman/italic characters … However, with the exception of a few common or short words, most of the words don't appear in the corpora.[13]
>
> VR: … apply this algorithmic sequence to the remaining words: 1. Latin. 2. Latin abbreviation. 3. Latinized Greek. 4. Galician-Portuguese. 5. Arabic.

The moral of this story is perhaps that the Voynich manuscript permits, indeed invites, radically different approaches to its interpretation. This author has taken the view that an analytical approach, inspired by Occam's razor and with strictly limited degrees of freedom, can yield a pathway toward a solution. The correspondence summarized above shows that an instinctive and heuristic approach is an alternative, which this author would not wish to dismiss. However, there are further avenues for an analytical solution, which we shall explore in later chapters.

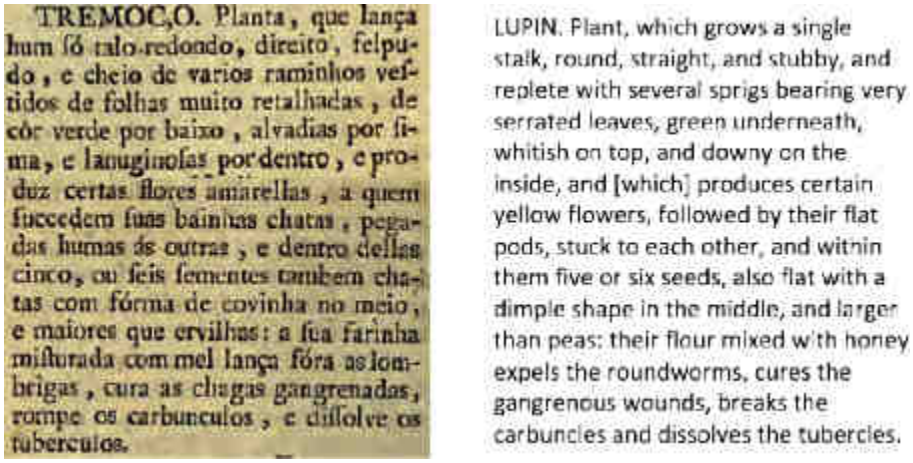
POSTSCRIPT: TREMOÇO

Following our correspondence on the presumed small-flowered lupin, as narrated above, and the introduction of medieval Galician-Portuguese as a possible underlying language, this author was impelled to search at least for the Galician word for lupin.

In modern Galician and Portuguese, and also in medieval Galician-Portuguese, the word for the lupin bean is tremoço.

In an eighteenth-century Portuguese dictionary of plants, animals and other useful organic and inorganic things (of which we reproduce an extract

in Figure 9.5), we found the description of the lupin, which appears in Figure 5.[14]



TREMOÇO. Planta, que lança hum só talo redondo, direito, felpudo, e cheio de varios raminhos veſtidos de folhas muito retalhadas, de côr verde por baixo, alvadias por ſima, e lanuginoſas por dentro, e produz certas flores amarellas, a quem ſuccedem ſuas bainhas chatas, pegadas humas ás outras, e dentro dellas cinco, ou ſeis ſementes tambem chatas com fórma de covinha no meio, e maiores que ervilhas: a ſua farinha miſturada com mel lança fóra as lombrigas, cura as chagas gangrenadas, rompe os carbunculos, e diſſolve os tuberculos.

LUPIN. Plant, which grows a single stalk, round, straight, and stubby, and replete with several sprigs bearing very serrated leaves, green underneath, whitish on top, and downy on the inside, and [which] produces certain yellow flowers, followed by their flat pods, stuck to each other, and within them five or six seeds, also flat with a dimple shape in the middle, and larger than peas: their flour mixed with honey expels the roundworms, cures the gangrenous wounds, breaks the carbuncles and dissolves the tubercles.

Figure 9.5. A description of the lupin, from José Monteiro De Carvalho, *Diccionario Portuguez Das Plantas, Arbustos etc*. (Lisbon, 1765), pp. 563, 564. *Public domain*

The lupin, then, was certainly known as a medicinal plant in the eighteenth century, and we might reasonably suppose that it was so known to the authors of the Voynich manuscript.

Could we then find, on page f006v which might illustrate the lupin, a glyph string approximating to the corresponding Galician word?

Among all the glyph strings, there was just one that had seven glyphs and seemed to have four consonants and three vowels in the right order. It was ﬞƈƚɑɋoϑϙ {k1aeo89}, shown in its context in Figure 9.6.

If this transliteration worked, it would give us a crib for seven of the most-common glyphs in the Voynich manuscript. We would have ﬞ = “t,” ƈƚ = “r,” ɑ = “e,” ɋ = “m,” o = “o,” ϑ = “ç” or “s,” and the final ϙ = final “o” or possibly a generic noun ending.

There is something encouraging about this crib; the glyphs are morphologically not too distant from the Latin letters with which we are trying to match them. We could imagine the ﬞ for example as an embellished “t.”

For the remaining glyphs on the page, we could refer to the glyph frequencies, with a view to matching common glyphs with common Latin letters.

For what it is worth, the “tremoço” crib, coupled with matching the frequencies of glyphs and Latin letters, yielded a provisional transliteration of page f6v into medieval Galician-Portuguese, as shown in Figure 9.7.

With the exception of the word “tremoço,” which we forced upon it, the transliteration does not seem to possess narrative or even poetic meaning, insofar as this author’s knowledge of Galician permits an interpretation. Nevertheless, the words seem to be pronounceable, to have a sensible balance between vowels and consonants, and to be capable of further refinement. We shall return to Galician-Portuguese in a later chapter.

Figure 9.6. Voynich manuscript: plant illustration on page f006v, with the glyph string that could possibly be transliterated to the Galician and Portuguese word “tremoço” (“lupin”). *Beinecke Rare Book and Manuscript Library*

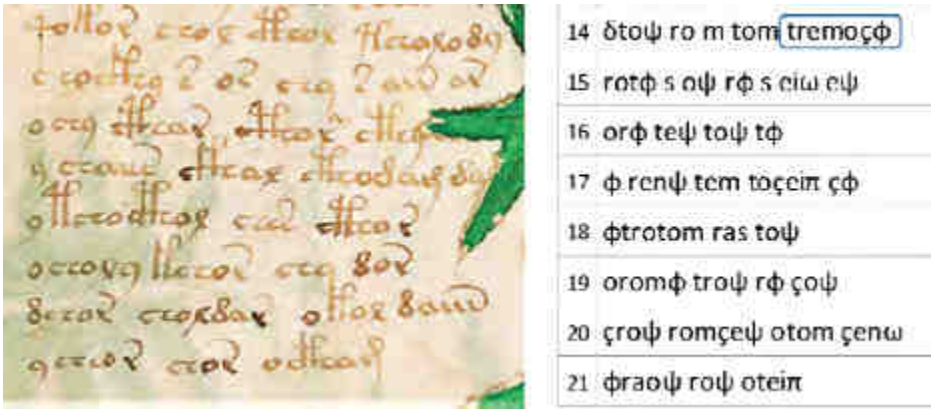

14 δtoψ ro m tom tremoçφ
15 rotφ s oψ rφ s ciω eψ
16 orφ teψ toψ tφ
17 φ renψ tem toçeiπ çφ
18 φtrotom ras toψ
19 oromφ troψ rφ çoψ
20 çroψ romçeψ otom çenω
21 φraoψ roψ oteiπ

Figure 9.7. Voynich manuscript, glyph strings on page f006v, and provisional transliteration to medieval Galician-Portuguese on the basis of crib from “tremoço.” Symbols φ, ψ, ω, π, δ represent “marker” glyphs which are assumed to have no meaning. *Beinecke Rare Book and Manuscript Library, graphics and transliteration by author*

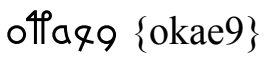
POSTSCRIPT: DEGREES OF FREEDOM

As we observed above, if we are free to select languages, transpositions, encipherments, anagrams and other mechanisms of our choice, we can probably make any preconceived narrative fit the Voynich manuscript. We are then at risk of falling into the Voynich equivalent of the Cooper Vortex.

By way of example: one proposal is that the Voynich glyphs were not written in an exotic or unknown language but were simply Italian letters written in an ornate script. In this view, the glyph {N}, which could be redrawn to resemble a "d," is in fact a "d"; the glyph {E} resembles a "c" and is in fact a "c"; and so on.[15]

This interpretation casts aside all reference to letter frequencies. It makes the letter "g," for example, correspond to {9}, the third most common glyph in the Voynich manuscript, which has a frequency of 10.9 percent, whereas in the OVI corpus of medieval Italian, "g" is the sixteenth most common letter, with a frequency of 1.9 percent. It is gas gif somegone had set out to gwrite Italigian with a battered typegwriter gwhich produced a "g," gmore often than gnot, no gmatter which key gyou hit.

A corollary of this proposal permits the researcher yet more degrees of freedom. The proposition is that the glyph strings, once retyped in conventional Latin letters, are not the final words that they represent. They are anagrams of those words. The following glyph string on f088r:

 {okae9}

is transliterated as "olaig," which is an anagram of "aglio," the Italian word for garlic.

Indeed, it is. However, there are 120 possible anagrams of "olaig"; if we select "aglio" and ignore the 119 others, we have given ourselves 120 degrees of freedom.

POSTSCRIPT: SENSE FROM NONSENSE

Here is a sequence of numbers between 1 and 26, produced by the random number function in Excel.[16] There was no human intervention in these numbers (except that they are arranged in strings of five). They are random, or if we prefer, nonsense:

2, 11, 22, 16, 14 16, 16, 12, 22, 12 17, 13, 6, 16, 9 4, 14, 16, 19, 5

We may now transliterate these numbers to Latin letters, using the mapping 1=A, 2=B, 3=C, etc.:

BKVPN PPLVL QMFPI DNPSE

Now we look for anagrams. The first three strings cannot produce a word in any language. But the fourth string is an anagram of “SPEND,” which is a real English word.

With 120 degrees of freedom, we were able to make sense from nonsense.

CHAPTER 10

ARIES

Having practically drawn a blank in our search for plant names in the "herbal" pages, we turned to the twelve "zodiac" pages. Here, surely, we had reasons to be cheerful. Unlike the plant pages, the subjects of the "zodiac" pages are unmistakable; they are clearly inspired by the constellations that are familiar to us from astrology. Every "zodiac" page has a central element in the form of an illustration, which, in most cases, conforms to our modern image of the zodiac sign: a pair of fishes, a ram, a bull, an archer, and so on.

Only two of these illustrations seem out of line with the modern perception: that for "Cancer" should be a crab but looks more like a pair of lobsters, while the representation of "Scorpio" has four legs instead of eight, has no sting, and looks like an athletic lizard.

In case there were any doubt about the intended theme of these pages, each page is annotated with a label in Latin script, in a language which appears to be Middle French, which undoubtedly conveys the name of a month. Furthermore, the month is in every case that which we associate with the corresponding sign of the zodiac. For example, the page for Pisces,

which in astrology is linked to the period February 19 to March 20, is annotated with a single word which seems to read “mars” (meaning “March” in Old French and even in modern French). The annotation is always below the central image.

We do not know whether the annotations were inserted by the original scribes of the Voynich manuscript. If they were, the scribes either understood French or knew how to write in a Latin script. Alternatively, the annotations could have been added by a Frenchspeaking owner of the book, at some unknowable time during the centuries between the fifteenth and the twentieth. There is plenty of space within the central illustrations, for an owner or reader to insert notes. The ink of the annotations looks similar in color to that of the glyphs, but is slightly darker, which perhaps indicates a more recent date.

The central elements of the “zodiac” pages, including their annotations, are summarized in Figure 10.1.

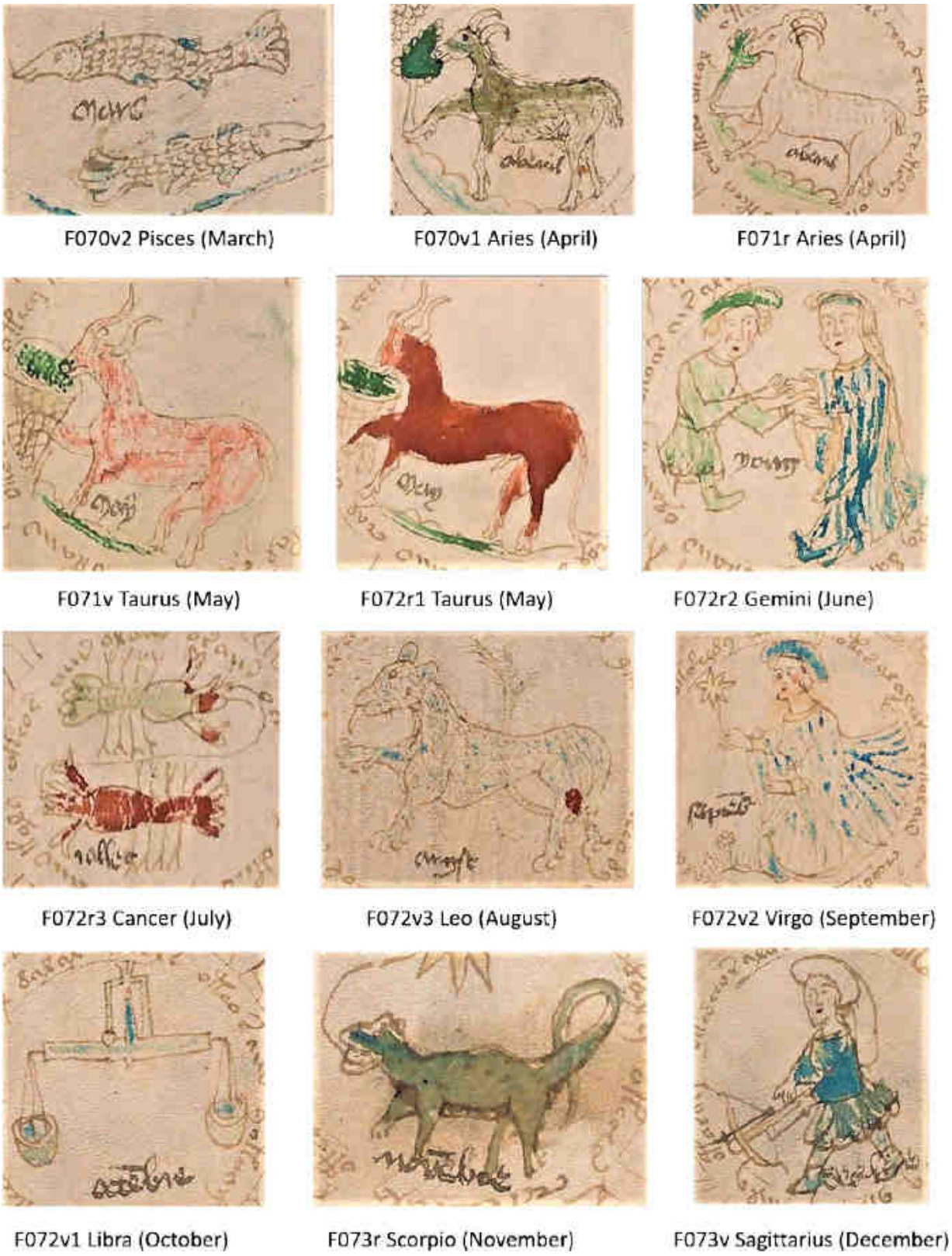


Figure 10.1. Voynich manuscript, extracts from the twelve pages of the “zodiac” section. Each extract shows the central “logo” and the accompanying annotation. *Beinecke Rare*

For unknown reasons, although there are twelve pages devoted to the zodiac signs, only ten signs are represented. The signs of Aries and Taurus have two pages each, while the signs of Capricorn and Aquarius, corresponding respectively to January and February, are missing. The two pages for Aries are very similar to each other; likewise, the two pages for Taurus.

The "zodiac" pages, missing two signs and duplicating two others, are clearly useless as a practical manual of astrology. This suggests that either the scribes did not know what they were writing or drawing, or that it did not matter. Whatever the owners or readers of the manuscript were expected to do, it was not to read their horoscopes or to foretell their futures.

Every one of the central illustrations is surrounded by two or three concentric rings of glyphs, arranged in strings that look like words. We have no way of knowing where the text is supposed to begin, or indeed whether it is continuous text with an end and a beginning. Within the rings of glyphs there are drawings of naked women, also arranged in concentric rings. Each woman is apparently bathing in a tub of water, recalling the women in the "balneological" pages. Each woman is flanked by something that looks like a flower with seven petals, or a seven-pointed star (and it is always seven), on the end of a wand or wire of some kind. Sometimes the woman holds the wand in her hand, and sometimes the wand is attached to the tub. Each of the women is accompanied by a string of glyphs that looks like a label, maybe referring to the woman, or maybe referring to the star.

Surely, we may hope that in each of the "zodiac" pages, we may find the name of a star from the respective constellation. And unlike in the plant pages, we do not have to guess what we are looking for. If the page refers to the zodiac sign of Pisces, we are looking for a star in the constellation of Pisces.

FLAGS OR KEYWORDS

As with our search among the herbal pages, we preceded our search of the zodiac pages with an investigation of possible keywords. From Table 7.6,

we had identified seven "words" that were common in the zodiac section but rarely occurred anywhere else: namely {okcos}, {okam}, {okay}, {okcc9}, {okco}, {okco89} and {ok9}. Unlike in the herbal section, most of these "words" were long enough that we could suspect them to be names of zodiacal concepts. It seemed a reasonable hypothesis that these "words" might recur in the astronomical or cosmological sections, but rarely or not at all in the herbal or biological sections.

In Table 10.1 is our first test of this hypothesis.

Table 10.1. Voynich manuscript, possible keywords for the astrology/zodiac section

"word"	v101	v104	"word" count, by section				
			astrology/zodiac	astronomical	cosmological	herbal	balne
	okcos	okcos	14	2	3	3	
	okco	okco	10	3	4	1	
	okco89	okco8Φ	9	0	5	2	
	ohcos	ohcos	6	2	2	2	

Source: author's analysis.

Here we see four "words"— {okcos}, {okco}, {okco89} and {ohcos}—which are common in the zodiac section (at least, relatively to the length of this rather short section); occur occasionally in the other "cosmic" sections and rarely in the herbal section; and are absent from the biological section. As in the herbal section, these "words" have affinities, either the initial {okc} or the final {cos}. So here are four candidates for flags or keywords that mark the zodiac section.

With that, we had a reason to start our hopeful search for the names of stars or constellations.

ARIES

To give ourselves two bites at the cherry, we started our search in the two pages which represent the sign of Aries: f070v1 and f071r. Each of the pages for Aries has a central illustration depicting an unmistakable ram. That on page f070v1 is reproduced in Figure 10.2.

It may or may not be significant, but the Voynich illustrations of the ram, on both of the pages where they appear, do not greatly resemble the images of Aries that have come down to us since antiquity. The Voynich ram is standing, with the right foreleg raised, and is looking forward, in the direction of a small green bush or tree. The classical Aries is practically always depicted in a reclining position, with the forelegs and rear legs folded, and with the head turned so that the animal looks backward.

Prescott Currier did not identify the languages on folios f070 or f071 (possibly because the glyphs were written in concentric circles, rather than horizontal lines). However, Dr. Lisa Fagin Davis has proposed that both pages are in Currier's Language B.

Our first step was to see whether any string of glyphs occurred more than once on either of these pages, or better yet, occurred on both pages. For this purpose, we focused on strings of at least five glyphs; we expected a star name to have a length of at least five letters in Latin script.

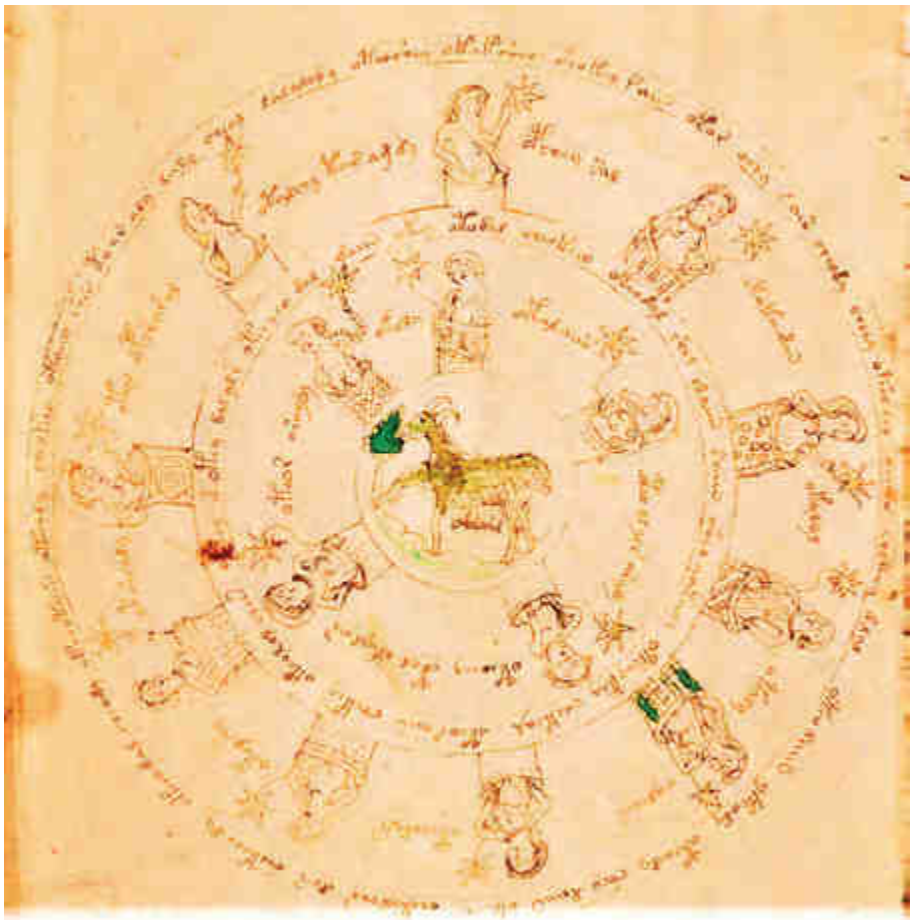


Figure 10.2. Voynich manuscript, page f70v1, depicting the zodiac sign of Aries. Color and contrast enhanced. *Beinecke Rare Book and Manuscript Library*

On pages f070v1 and f071, we found five such strings. The most promising was the string of seven glyphs:

ⱺⱪⱁⵒⵒⱪⱁ [1cohcco]

which occurred on page f070v1 in the second ring, and in a slightly different form, with one glyph duplicated and what looked like an accent over the first glyph:

ⱺⱪⱪⱁⵒⵒⱪⱁ [1ĉohcco]

on page f071r in the outer ring.

The second step, on the assumption that both pages were in Language B, was to attempt transliterations from v104 to European languages (plus Hebrew), on the basis of letter frequencies, as per the "Gold Bug" approach. In some cases, we tried variants with and without accents, since combining, for example, an "e" and an "ë" changes the letter frequencies.

The third step was to see whether better transliterations would result from mapping glyphs to letters according to their position (initial, internal, or final); or alternatively from a position-dependent mapping of glyph pairs to digraphs (which we could do for medieval Italian, medieval Latin, and Old French).

Table 10.2. Voynich manuscript, "zodiac" section, pages f70v1 and f71r (Language B), v104 transliteration: mapping of selected "words" to selected European languages

"word"	glyph string	⟨ccolkcco	ccolkcco	olkcos9	olkcc9	olkos9
	alternative[1]	⟨ccolkcco	ccolkcco		olkcc9	
target language	RSQ[3]	mappings based on glyph and letter frequencies[2]				
Albanian	98.4%	Ë'JINTTI Ë'JINJI	ËTINTTI ËTINJI	IOTIRE	IOTTE IOJE	INIRE
Finnish	97.8%	N'JILTTI N'JILJI	NTILTTI NTILJI	IKTISA	IKTTA IKJA	ILISA
Welsh	97.5%	E'TARYYA E'TARTA	EYARYYA EYARTA	AWYAID	AWYYD AWTD	ARAID
Latin, classical	97.3%	T'QUERAAE T'QUERQUE	TAERAAE TAERQUE	EMAEUI	EMAAI EMQUI	EREUI
Italian, modern[4]	97.2%	O'GARIIA O'GARGA	OIARIIA OIARGA	ACIATE	ACIIE ACGE	ARATE

Source: author's analysis. Notes: (1) alternative based on cc => cc;

(2) marker glyphs were transliterated; (3) denotes correlation coefficient with Voynich language B; (4) denotes variant without accents. Letter frequencies by courtesy of Stefan Trost Media.

The results are shown in Table 10.2.

What are we to make of this?

ALPHABET SOUP

At first sight, the transliterations look an alphabet soup. What we can say is as follows:

- In the five target languages that we tried (each with correlation of over 97 percent with Language B), the mapping yielded text strings which were generally pronounceable.
- On checking the mapped text strings against corpora of the respective languages, most of the strings were not found.
- The one text string that we found in the corpora was ARATE, which is a real word in modern Italian and also in the OVI corpus of medieval Italian. This word is infrequent and is a past participle of the Italian verb arare (to plough).

STAR NAMES

However, our task has only just begun. We now have to examine whether any of the pronounceable transliterations could conceivably be the name of a star in the constellation of Aries.

In 2016 the International Astronomical Union (IAU), after much deliberation and with some reluctance (apparently not wishing to get involved in popular culture) set up a Working Group on Star Names. Through the efforts of this group, the IAU has been publishing a "List of IAU-approved Star Names."[1] As of January 1, 2021, there were 449 stars in the list. In the constellation of Aries, as of that date, there were six stars in the list, as follows:

Bharani	(HR 838 or 41 Arietis)
Botein	(HR 951 or δ Arietis)
Hamal	(HR 617 or α Arietis)
Lilii Borea	(HR 824 or 39 Arietis)

Mesarthim	(HR 546 or γ1 Arietis)
Sheratan	(HR 553 or β Arietis).

The star Lilii Borea was first named in 1757 by Nicolas-Louis de Lacaille,[2] who considered it part of Lilium (the Lily), a sky zone that is no longer recognized as a constellation. If we presume, as we have so far, that the Voynich manuscript was written before 1757, then the producers and their scribes cannot have known this name, and we do not need to search for it.

Bharani was a concept in Indian astronomy from around 1500 BC onward; the name applied to the three stars now known as 35, 39, and 41 Arietis. It is conceivable that the producers of the Voynich manuscript knew of this name.

The other four-star names (Botein, Hamal, Mesarthim, and Sheratan) are of Arabic origin, and have been used since antiquity. Here it is surely a reasonable assumption that these four names were familiar to the producers of the Voynich manuscript.

Can we then find on the Voynich pages f070v1 and f071r, glyph strings corresponding to Bharani, Botein, Hamal, Mesarthim, or Sheratan?

The longest of these names in Latin script (Mesarthim) would probably have been written in Arabic المسارثم (al-musarthim), meaning “the tracker.” The Arabic word has eight letters including the definite article ال; it is unlikely that in written Arabic, the definite article would have been omitted. The third letter (reading from right to left) and the eighth are both م. From the glyph strings that we selected, our only possible candidate for this name is ctcco⁋cco [1ĉohcco]; but whether we read from left to right or from right to left, we cannot make this string match المسارثم.

The name Sheratan, in Arabic, would have been written الشراطان (aš-šarāṭān), or “the two signs.” Here the Arabic word has eight letters, of which the first, fifth and seventh are ا {a}. Again, we have no eight-glyph string in which the first, fifth, and seventh are the same.

Botein would have been written البُطَين (al-butayn) or “the little belly.” The Arabic word has six letters with no repetitions. Among the glyph strings that we selected, there is none that can be made to correspond to this word.

Hamal would be رأس الحمل (ra's al-hamal, "head of the ram") or simply الحمل (al-hamal, "the ram"). If we are searching for a string which represents الحمل, we need a string of five glyphs with no repetitions. We have two candidate strings of five glyphs: o𐋏c𐋏c9 [okcc9] and o𐋏o8𐋏9 [oho8Φ]; but each of them has a repeated glyph.

Finally, Bharani would have been written भरणी in Sanskrit (or in modern Hindi). This word has four letters—भ {bh}, र {r}, ण{n}, ी {i}—with no repetitions. On pages f070v1 and f071r, there are two glyph strings which are possible candidates: o𐋏aq [okae] and o𐋏c9 [ohcΦ], or o𐋏c𐋏9 [ohĉΦ] allowing for mis spelling. But for these glyphs to spell "Bharani," o {o} the most common glyph of all, would have to correspond improbably to भ {bh}, the thirty-eighth most frequent glyph in Hindi.

As far as we can determine, therefore, none of the star names in Aries, as currently prescribed by the IAU, occurs in either of the pages for Aries in the Voynich manuscript.

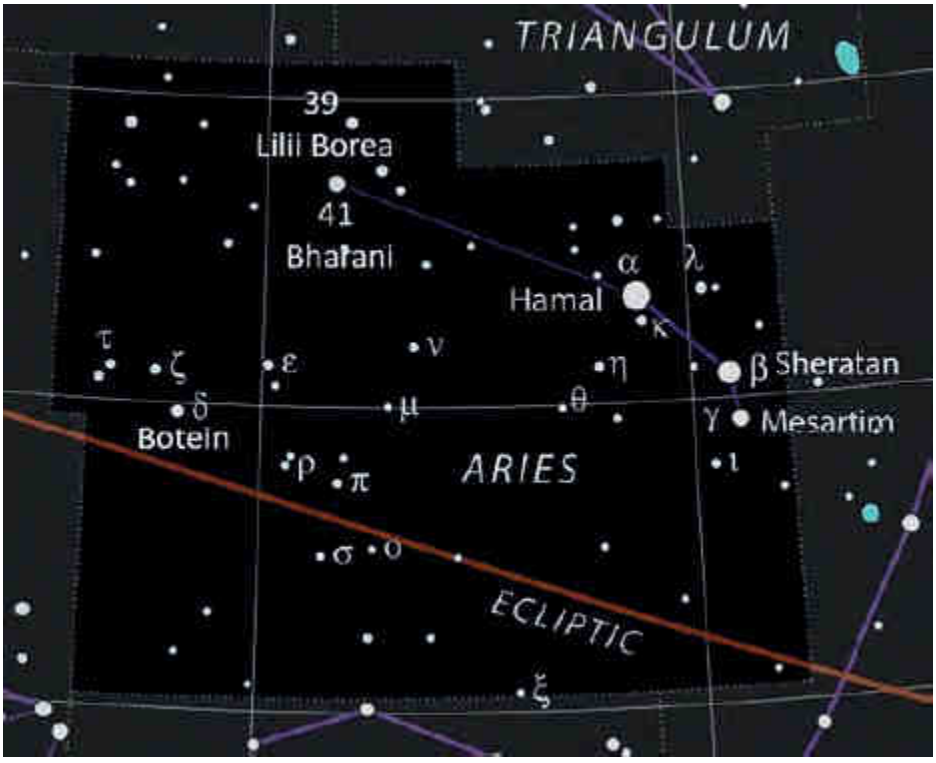


Figure 10.3. A modern map of the constellation of Aries. *International Astronomical Union, annotations and graphics by author*

STAR CATALOGS

However, we are not obliged to assume that the star names that are accepted in modern times are those that were known to the producers of the Voynich manuscript.

Catalogs of stars have existed since antiquity. The authors of the Voynich would surely have been familiar, for example, with the *Almagest*[3] of the Egyptian mathematician and scientist Claudius Ptolemy (ca. 100–ca.170 AD). The *Almagest*, completed sometime after 150 AD, is a treatise on the apparent motion of the observable stars and the known planets. It includes a catalog of 1,029 fixed stars.

Ptolemy recognized Aries as a constellation, and his *Almagest* listed thirteen stars in Aries. However, he did not attribute any proper names to them. For example, the star that we know today as Hamal or α Arietis, Ptolemy described as "the star under the knee-bend." For each of the stars that he allocated to Aries, he recorded the longitude, latitude, and magnitude (apparent brightness), but nothing more. So, if the Voynich authors used the names of stars, they could not have come from Ptolemy.

The most contemporary star catalog, for the authors of the Voynich, would have been that of Muḥammad Ṭaraghāy ibn Shāhrukh ibn Tīmūr (1394–1449), more commonly known as Ulugh Beg ("Great Prince"), the Turkish governor of Transoxiana and Turkestan. His astronomical handbook, first published in 1437, is variously known as *Zīj-i Ulugh Beg, Zīj-i Jadīd-i Sulṭānī*, or *Zīj-i Gūrgānī*. It listed 1,018 stars, eleven fewer than those in the catalog of Ptolemy.[4] Again, Ulugh Beg recorded the positions and magnitudes of the stars, in most cases from original observations, but he did not list their names.

If the authors of the Voynich wanted to look up the names of stars, a possible reference would have been *The Book of Fixed Stars*, or in Arabic كتاب صور الكواكب (kitāb suwar al-kawākib) by Abd al-Rahman al-Sufi, written around 964 AD.

Here, finally, they would have found charts of the constellations, annotated with the names of some of the principal stars. Al-Sufi's chart of Aries is reproduced in Figure 10.4.

Figure 10.4. Depiction of the constellation of Aries and its principal stars, from *The Book of Fixed Stars* by Abd-al-rahman al-Sufi. *Public domain*

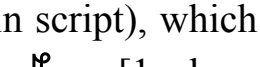
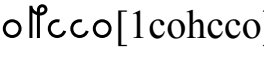
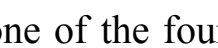
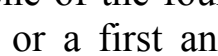
If they could read the Arabic script, they would have noted that the title of the chart was كوكبه الحمل على عابوب ف السما (kawkabat al-hamal 'ala 'abub fi al-sama, or "the constellation of the ram on the mouth[?] of the sky") and they would have found the names of two stars: الطبين (al-butayn) and الشرطين (al-sharatayn), but not the word الحمل (al-hamal) as the name of a star. Every other star in Aries was annotated with a single Arabic letter but not a name. In short, they would found some of the star names which are in use today, and for which we have searched unsuccessfully in the respective pages of the Voynich manuscript.

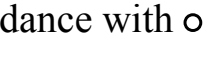
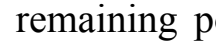
OTHER STAR NAMES

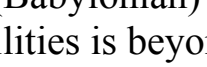
If we cannot find the Sanskrit/Hindi or Arabic names of stars in the Aries pages, it still remains to consider whether the producers might have inserted star names in other languages.

This author has identified the following alternative names for the principal stars of Aries, in various languages, for which we are indebted principally to Richard Hinckley Allen's classic *Star Names: Their Lore and Meaning*:[5]

- α Arietis: Aloros (Akkadian); Ailuv (Assyrian); Anuv, As-kar, Dil-kar, Dil-gan, Ku, Lu-lim, Lu-nit, Si-mal, Si-mul (Babylonian); Ayil (Hebrew); Cornu (Latin).
- β Arietis: Sartai or Scartai (attributed to Hebrew); Açvināu (Sanskrit/Hindi); Borealior Signorum (Latin);
- β and γ Arietis (together): Marhū-sha-rishu-ku (Babylonian); Pikutorion (Coptic); Padevar (Persian); Bashish (Sogdian)
- γ Arietis: Mesartun (Arabic?); M$^{\varepsilon}$shār$^{\varepsilon}$tīm (Hebrew); Açvayujāu (Sanskrit/Hindi)
- δ Arietis: Nir Al Buṭain (Arabic); Lucida Ventris (Latin)

None of these names has a repeated letter (at least not when expressed in Latin script), which tends to rule out a mapping to  [1ĉohcco],  [1cohcco],  [okccΦ], or  [ohĉΦ].

None of the four-letter or five-letter names of stars has a first and third letter, or a first and fourth letter, in common, which tends to rule out a concordance with  [oho8Φ] or  [okco].

The remaining possibilities, if we confine ourselves to repeated glyph strings, are a concordance with [okae], or  [ohcΦ], for which Anuv (Babylonian) or Ayil (Hebrew) might be candidates. To explore these possibilities is beyond the scope of this book.

In any case, Occam's razor obliges us to suppose that the producers and scribes of the Voynich manuscript, if they wished to mention specific stars

at all, would have chosen names that were current in their time. To consider that they would have used more ancient names requires complex assumptions, and we wish to follow the path of simplicity.

TAURUS

Let us turn then to the other zodiac sign which has two pages in the Voynich manuscript: Taurus.

Here we have not just the dedicated pages f071v and f072r1, but also a third resource in the form of page f068r3. This page, typically considered one of the "astronomical" pages, includes a drawing of a group of seven objects that might be stars. Six of the "stars" are drawn with six points; the seventh, and rightmost in the drawing, has seven points. Many researchers have suggested that this drawing could represent the Pleiades, the well-known group of seven stars in the constellation of Taurus. There is a large star nearby, to the left and slightly below the cluster; it is in the right place to be the bright star Aldebaran. The drawing is reproduced in Figure 10.5. In Figure 10.6, for comparison, is a modern map of Taurus.

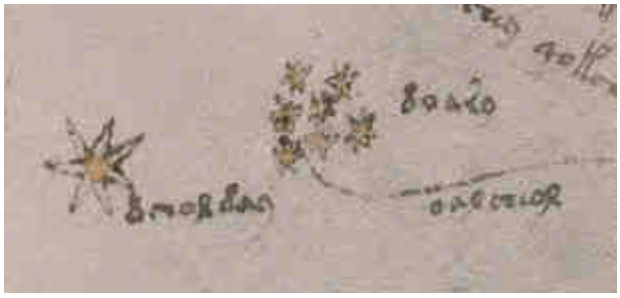


Figure 10.5. Voynich manuscript, extract from page f068r3, showing a group of seven "stars." *Beinecke Rare Book and Manuscript Library*

Let us cut to the chase and look up all the modern names of stars in Taurus. The IAU has generously given official names to sixteen stars in Taurus, as follows:

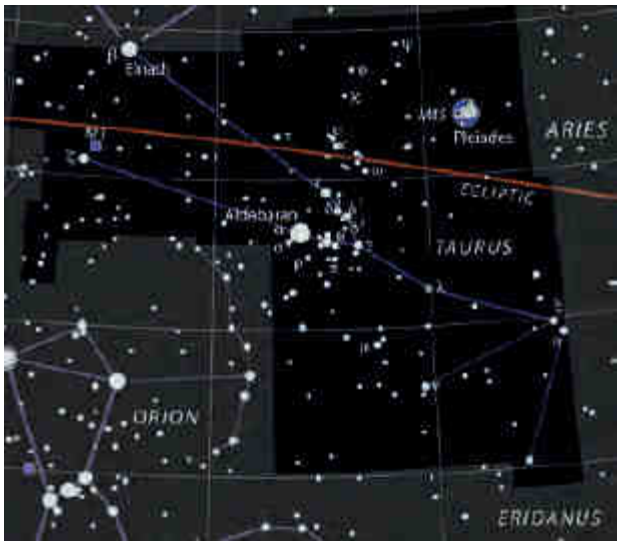


Figure 10.6. A modern map of the constellation of Taurus. *International Astronomical Union*

Ain	(HR 1409) ε Tauri
Alcyone	(HR 1165) η Tauri
Aldebaran	(HR 1457) α Tauri
Asterope	(HR 1151) 21 Tauri
Atlas	(HR 1178) 27 Tauri
Celaeno	(HR 1140) 16 Tauri
Chamukuy	(HR 1412) θ2 Tauri
Electra	(HR 1142 17 Tauri
Elnath	(HR 1791 β Tauri
Hoggar	(HD 28678) _ Tauri
Maia	(HR 1149) 20 Tauri
Merope	(HR 1156) 23 Tauri
Pleione	(HR 1180) 28 Tauri
Prima Hyadum	(HR 1346) γ Tauri
Secunda Hyadum	(HR 1373) δ Tauri
Taygeta	(HR 1145) 19 Tauri

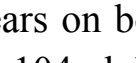
This list includes Aldebaran or α Tauri, one of the brightest stars in the sky. This star was known as Rohini in Indian astronomy and as Lampadias

by the Greeks of antiquity. Richard Hinckley Allen, in *Star Names, Their Lore and Meaning*, citing earlier authors, listed the following variants and alternatives:

> Addebiris, Debiron (Schickard)
> Aldebara, Aldebaram, Oculus australis (Riccioli)
> Aldebaron (Costard)
> Aldeberan (Chaucer)
> Aldeboran (Spenser)
> Al Fanīḳ, Al Fatīḳ, Al Muḥdij (Al Bīrūnī)
> ʻAin al Thaur (Arabic), Atin and Hain Altor
> Bull's Eye (early English)
> Hrusa (Bohemian)
> Left Eye (Aben Ezra)
> Oculus Tauri (Latin)
> Sāïḳ al Thurayya (Niebuhr)
> Tāli al Najm, Hādī al Najm (Arabian sources)[6]

And we may add that Aldebaran derives from the Arabic الدبران (al dabarān).

The IAU has listed all seven stars of the Pleiades—Alcyone, Asterope, Celaeno, Electra, Maia, Merope, and Taygeta—the seven daughters of Atlas and Pleione, as well as their parents, with their traditional names. For the Pleiades as a group, Hinckley listed a number of traditional European names including Aften Hoehne (Danish), Clock Henne (English), Crannarain (Gaelic), Gallinella (Italian), Gallinelle, Pulsiniere, Poussinière (French), Gluck Henne (German), Наседка (Russian), Πούλια or Πούλεια (Greek).[7] We shall look for these names on the two Taurus pages as well as on page f068r3 of the manuscript.

Pages f071v and f072r1 represent the zodiac sign of Taurus. On these two pages, there is only one string which has at least four glyphs and appears on both pages. That string consists of the glyphs  , which in our v104 alphabet is [okĉΦ]. We already found several occurrences of this

glyph in the pages for Aries, so we do not expect it to represent a star in Taurus. In any case, our transliterations of this glyph string did not yield a recognizable word in any of the languages that we tested.

On page f068r3, there are three glyph strings in the neighborhood of the seven-star cluster.

- The string ⟨8c1oe8aΦ⟩, adjacent to the nearby large star, could be a single "word" [81oe8aΦ] but more probably is two "words"; the backward stroke of the ⟨e⟩ [e] gives us a clue that we might be looking at a word break. If so, the first "word" is ⟨8c1oe⟩ [81oe], which occurs seventeen times in the Language A pages but only four times in the B pages; and the second "word" is ⟨8aΦ⟩ [8aΦ], which occurs only once in the A pages and never in the B pages. We would like to think that these two words are a label for the star. Could it be Aldebaran?
- The string ⟨8oaΨΦ⟩ [8oaΨΦ] could be a label for the star cluster itself.
- The string ⟨oae1coe⟩ [oae1coe] is adjacent to the curved line which links the star cluster to the sun. Again, we suspect that the ⟨e⟩ [e] denotes a word break and that the string is probably two "words." If this string has anything to do with Taurus, its relevance must lie in the first word ⟨oae⟩ [oae].

We can force the putative star name ⟨8c1oe8aΦ⟩ [81oe8aΦ] to read Aldebaran, but only if we apply a crib on the following basis:

⟨8⟩ [8] = AL or AR

⟨c1⟩ [1] = D

⟨o⟩ [o] = E

⟨Ψ⟩ [Ψ] = B

⟨a⟩ [a] = A

⟨Φ⟩ [Φ] = N

In that case the string ȣoaʔ9, which we have been guessing is a label for the seven-star cluster, would read ALEA?N or AREA?N, and the string oaɣ ƈƈcoɣ [oae 1coe] would read EAB D?EB, none of which makes much sense.

If we were to insist on a one-to-one mapping of glyphs to letters, none of these strings could correspond to the modern names of Aldebaran or any of the stars of the Pleiades cluster. The strings are either the wrong length or have repeated glyphs in the wrong places.

We therefore had to fall back on the "Gold Bug" approach and to see whether any of these strings could be transliterated into sensible words in any European language. For the three pages f068, f071, and f072, Dr. Lisa Fagin Davis has proposed that the text is in Language B. We tested glyph-to-letter mappings, with and without position-dependency. We could not test mappings of glyph pairs to digraphs, as three of the five glyph strings had odd numbers of glyphs, and there was no sensible way to parse them.

The results are summarized in Table 10.3.

Again, we found that several European languages yielded transliterations that were pronounceable, in the sense of having a reasonable concordance of vowels and consonants. In particular the transliteration from Language B to Romanian yielded the intriguing result "riă neră," which means something like "black river" (but in Italian). However, none of the transliterations approximated to any of the modern names of the principal stars in Taurus, nor to any of the historical names of Aldebaran or of the stars in the Pleiades cluster, at least not those that we had found.

Table 10.3. Voynich manuscript, "astronomy" section, page f068r3 (Language B): mappings of selected "words"

"word" and location	glyphs	ȣƈƈoɣ ȣa9	ȣoaʔ9	oaɣ ƈƈoɣ
	v101	81oe 8a9	8oay9	oae 1coe
	v104	81oe 8aΦ	8oaΨΦ	oae 1coe
	position	f068r3, label	f068r3, label	f068r3, label
target language	correlation with B	mappings based on letter frequencies[2]		
Albanian	96.9%	ENTR EIΦ	ETIΨΦ	TIR NETR
Welsh	96.9%	ERYI EAΦ	EYAΨΦ	YAI RDYI

Italian	96.2%	ORIT OAΦ	OIAΨΦ	IAT REIT
Romanian	96.0%	ANRĂ AIΦ	ARIΨΦ	RIĂ NERĂ
Finnish	95.1%	NLTS NIΦ	NTIΨΦ	TIS LATS
target language	correlation with B	mappings based on frequencies of initial, internal and final letters		
Italian, medieval[1]	dependent on position	AOTA ARE	ATRΨE	DRA SETA
Latin, medieval		ATAS AIT	AAIΨT	EIS IUAS
Old French		CUNS CIT	CNIΨT	EIS SENS

Source: author's analysis. (1) denotes variant without accents (2) Marker glyphs are not transliterated. Letter frequencies in modern languages by courtesy of Stefan Trost Media; in medieval English from Moreno, Marsha Lynn (2005); in medieval Italian from author's analysis based on OVI corpus.

OR ELSE …

Otherwise, we must consider the possibility that the Voynich pages on the signs of the zodiac do not contain the name of any star in the respective constellations, in any language. We could easily imagine that the pages were simply intended to be a kind of manual, albeit incomplete and with redundancies, in which the reader's birth sign was linked to various homilies or remedies.

If so, our search for star names has been a wild-goose chase.

CHAPTER 11

CAPPELLI

Figure 11.1. Adriano Cappelli. *Public domain*

In this chapter, we shall explore the possibility that the Voynich manuscript was not enciphered, nor translated or transliterated from an unknown language, but was written in the language that was ubiquitous in European texts during the medieval era: namely, Latin.

To the foreseeable objection that the glyphs, with a few exceptions, do not seem to resemble Latin letters, we have this rejoinder. Throughout the Middle Ages, scribes wrote on parchment or vellum: materials which were derived from animal skins and which therefore were expensive. In order not to waste valuable material, the scribes often wrote in an abbreviated Latin. As we shall see, the abbreviations were often radical.

For an understanding of Latin abbreviations, our primary source is the classic work of Adriano Cappelli, *Lexicon Abbreviaturarum: Dizionario di Abbreviatture Latine ed Italiane* [Lexicon of Abbreviations: Dictionary of Latin and Italian Abbreviations], first published in 1899. The edition to which we have referred is that of 1929,[1] which is available online.

Cappelli defined six mechanisms for the abbreviation of Latin words, as follows:

- by truncation, in which the end of the word was replaced by a sign, for example the Latin word OMNIS (“all”) could be written
- by contraction, in which selected letters of the word were removed, for example, IAMSCRIPTIS (“the above mentioned”) could be written ISTIS
- by abbreviation signs with their own meaning, for example an initial or could be equivalent to CO-, COM-, CON-, CUM-, or CUN- (generally giving the sense of “with” or “together”), and a final superscript or could be equivalent to -IS, -OS, -S, or -US (generally signifying the plural); noting the resemblances to the ubiquitous Voynich glyph {9} and the common glyphs {y} and {s}, we shall return to these signs in particular
- by abbreviation signs with a meaning relative to the context; for example which, with horizontal lines appended, could variously mean QUAE, QUAM, QUE, QUI, QUO or QUONDAM (all of which were variants of “what” or “which” or “who”)
- by superscripts; for example, a superscribed could stand for the suffixes “-MUM or -TUM; and here we note the resemblance to the Voynich glyph {m}
- by conventional signs, for example & = ET (“and”), as well as Roman and Arabic numerals.[2]

However, *Lexicon Abbreviaturarum* is more than a compilation of abbreviations. Cappelli allocated an introductory section of seventy-three pages to the concepts of Latin abbreviation, including several facsimiles of medieval Latin texts in which abbreviations were pervasive (such as the San Savino document which we mentioned in Chapter 3). The bulk of his book, occupying 428 pages, was devoted to thousands of examples of handwritten Latin words or phrases which included some form of abbreviation. This

section is a gold mine, not only of the abbreviations in all their profusion, but also of ordinary Latin letters written in their full phonetic form. These examples demonstrate the extraordinary diversity of Latin handwriting in the medieval era.

Cappelli's examples of Latin handwriting reveal that many of the most common Voynich glyphs have some degree of morphological resemblance either to an ordinary Latin letter or to a Latin abbreviation. This allows us to pursue the hypothesis that some kind of mapping between glyphs and Latin letters is possible, on the basis of visual resemblance rather than frequencies. If it is, we have two ways to approach the mapping: from Voynich to Latin, and from Latin to Voynich.

For the modern reader, the greatest omission in Cappelli's work is the absence of any clue to the frequency of the abbreviations, and indeed of any of the handwriting examples. That could only be expected of a work from the precomputer era. *Lexicon Abbreviaturarum* gives us no idea whether a given abbreviation is sufficiently common that we should be looking for it in the Voynich manuscript, or so rare that we could forget about it. In our research we endeavoured to rectify that omission.

In recognition of the inspiration that we drew from the work of Adriano Cappelli, perhaps we could call this methodology the "Cappelli approach."

VOYNICH TO LATIN

If the scribes of the Voynich manuscript were writing in some form of Latin, we can imagine two scenarios. The first is that they were copying from source documents which were already in abbreviated Latin, and they were instructed to reproduce the abbreviations as they saw them. Alternatively, they were copying from source documents which were not abbreviated, and they were instructed to abbreviate according to some standardized system (and the Voynich manuscript is highly standardized in its calligraphy). In either event, at least two phenomena would occur which would greatly impede what we earlier called the "Gold Bug" approach, or indeed any attempt at transliteration based on glyph and letter frequencies.

Firstly, the glyph frequencies would diverge from the letter frequencies in Latin as it is written in full. To take Cappelli's example of the Latin abbreviation , which stands for "generalissimum": the word contains some of the most common letters in the Latin alphabet, including A, E, I, L, S, and U; but we would search in vain for these letters in the abbreviation.

Secondly, the mapping from glyphs to Latin letters would not be one to one. We have seen above that the sign , as a prefix, can mean CO-, COM-, CON-, CUM-, or CUN-, and the very similar sign , as a suffix, can mean -IS, -OS, -S, or -US. If the Voynich scribes combined these two signs into their glyph {9}, or if the various modern transliterations of the Voynich manuscript simply assumed that all variants of this glyph were the same, then we have to interpret the glyph {9} according to its position in the Voynich "word" and according to the context provided by the rest of the "word."

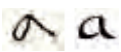
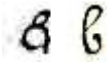
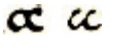
In Tables 11.1A and 11.1B, we have assembled some of the most common glyphs in the Voynich manuscript; from Cappelli's work, and we have selected medieval Latin letters or abbreviation signs which seemed to be visually similar.

In many cases, for a given glyph there are several visually similar counterparts in Latin. In such cases it is instructive to look at the frequency of the possible Latin letters or expressions. As a reference for this purpose, we used part of the LatinISE corpus,[3] specifically the 1,000 most frequent words in the medieval Latin subcorpus, and calculated the number of occurrences of each letter or expression. This body of text, by our estimates, contained 1,313,126 words and 5,441,442 letters. The average word length was 4.14 letters, which was short by the standards of classical Latin (7.22 letters in the Packard Humanities corpus),[4] but quite comparable with the Voynich manuscript, where the average length of a "word" is 3.90 glyphs in v101 and 4.10 glyphs in v104.[5]

Following are some examples of how this mapping could be applied to transliterate the Voynich manuscript to medieval Latin.

Table 11.1A. Visual resemblances between common Voynich glyphs and Latin letters

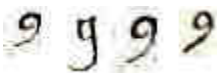
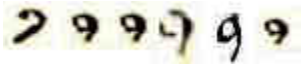
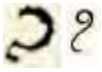
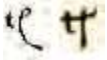
glyph	v101	v104	position	Latin manuscript forms	Latin letters

ο	o	o	anywhere		O
ϙ	9	Φ	anywhere		I U O Q G
ϙ	9	Φ	initial		E Q G
ɑ	a	a	anywhere		E A
ɑ	a	a	initial		AE
ϲ	c	c	anywhere		C
ϖ	1	1	anywhere		T CC
ϖ	1	1	initial		T
ϧ	e	e	anywhere		4
ϐ	8	8	anywhere		I S O
ϐ	8	8	initial		E B
ϐ	8	8	final		S
ⅈ	h	h	anywhere		L P F LL PP K
ⅈ	h	h	initial		P L K
ⅉ	k	h	anywhere		T SS
†	4	4	anywhere		4
ⅵ	m	îΩ	anywhere		M
ϑ̂	2	1’	anywhere		ET?
ϲϲ	C	ĉ	anywhere		A
ϲϲ	C	ĉ	initial		A CE
ϛ	s	s	anywhere		N
ϛ	s	s	initial		C

◌	1o	1o	anywhere		CA
◌	n	iN	anywhere		I A N II
◌	n	iN	final		N
◌	cc	cc	anywhere		CC
◌	p	iΠ	final		T
◌	*	Π	anywhere		X
◌	N	Ω	anywhere		I V

Source: author's analysis. Glyph counts from v101 transliterations. Latin letters and abbreviations from Cappelli (1923). Latin corpus counts from LatinISE corpus, medieval sub-corpus.

Table 11.1B. Visual resemblances between selected Voynich glyphs and Latin abbreviations

glyph	v101	v104	position	Latin abbreviation	Latin letters abbreviated
◌	9	Φ	initial		CON- CO- COM- CUN-
◌	9	Φ	final		-US -IS -S -NUS -QUE -OS
◌	y	Ψ	final		-TER -RE -RI
◌	m	îΩ	final		-M -TUM -MUM
◌	s	s	anywhere		ONCIA (weight)
◌	s	s	initial		CON- CUM- COM- CUN-
◌	n	iN	final		-NTES
◌	p	iΠ	final		-TIS
◌	*	Π	superscript		-ARIUS

ꝛ	*	Π	final		-ERE -IIS -IS -NIS -RIS -SIS
ꝰ	N	Ω	final		-ER -TUR -UR

Source: author's analysis; Latin abbreviations from Cappelli (1923)

The glyph o {o} is the most common in the Voynich manuscript, occurring 25,043 times. We have far too many occurrences of o {o}, an issue to which we shall have to return. Cappelli's examples suggest that the only plausible Latin counterpart is the letter O, written phonetically. In medieval Latin the letter O is rarely part of an abbreviation sign.

The glyph �o {9} is the second most common in the Voynich manuscript, occurring 17,436 times in v101. Of these occurrences:

- In 15,066 cases, the glyph is in the final position with regard to a Voynich "word." In these cases, the possible counterparts in *Lexicon Abbreviaturarum* include the abbreviation signs and , standing interchangeably for -IS, -OS, -S or -US; or , standing for -NUS; or , standing for -QUE. In selecting which of these counterparts is most probable, we can be guided by their frequencies in the LatinISE corpus, which suggest that "-us" is the most likely interpretation.
- In 1,619 cases, the glyph is in the initial position. The potential Cappelli counterparts are and , both standing for CO-, COM-, -CON, -CUM, or -CUN; also possibly (one variant of the initial E); and or (initial Q). Our strongest candidate for the initial �o {9} is the Latin "con-."
- In 325 cases, the glyph stands alone, looking like a single letter "word"; perhaps our best counterpart is the Latin word CUM ("with").
- In the remaining 436 cases, the glyph appears to be within a word, in which case Cappelli's examples suggest that it might transliterate to the Latin letter I or U. In these cases, there is always the possibility that there is not one word but two (or

more) crammed together, and the glyph is at the beginning or end of one of them.

Another glyph which must be interpreted in context is ⟑ {m} (if indeed it is a single glyph), which occurs 4,112 times in the v101 transliteration. In 4,071 of these instances, the glyph is at the end of a "word," in which case possible counterparts in *Lexicon Abbreviaturarum* are the abbreviation signs , standing for M-, or , standing for -MUM or -TUM.

Of the top twenty glyphs in the v101 transliteration, there are three which do not have obvious counterparts in *Lexicon Abbreviaturarum*:

- ⍜ {e), which occurs 10,653 times. In *Lexicon Abbreviaturarum* the only similar shapes are , and , which are variants of the Arabic numeral 4. Intriguingly, Cappelli listed a compound abbreviation , "4tis" or "quartis," meaning something like "to the fourths" or "to the quarters," and on page f116r we see an uncannily similar glyph string {ep}. But Cappelli lists no Latin letter, or an abbreviation for a letter, with a resemblance to the glyph.
- ⍭ {4}, which occurs 5,408 times. A similar shape appears in *Lexicon Abbreviaturarum* as , representing the Arabic numeral 4, but never as a Latin letter or abbreviation. However, in 5,265 cases the glyph ⍭ {4} is attached to an ⍥ {o}, with the resulting ⍭⍥ {4o} probably constituting a single glyph.
- ⍼ {2}, which occurs 3,359 times. The superscript or accent is footloose, and in variants of this glyph, appears sometimes to the left and sometimes to the right of the pedestal. We could hazard a guess that this "glyph" and its variants are actually two glyphs, namely ⍜ {9} and ⍼ {1}. In this case *Lexicon Abbreviaturarum* allows us the possibility of a counterpart in the two Latin letters E and T, which possibly could be written or ET (in English, "and"). An alternative idea (on which,

more later) is that the glyph, with its two inseparable components, is a representation of the Latin QU.

Apart from these three glyphs, we can work our way down the Voynich frequency table and in each case, advance reasonable hypotheses about interpretations of the glyphs as Latin letters or abbreviations.

However, if these hypotheses are to have any validity, at a minimum we need to examine, say, the ten most common "words" in the Voynich manuscript and see whether any of them map to common words in medieval Latin.

For example, the two most common "words," both occurring 739 times in the v101 transliteration, are ꝺɑ ııꝰ {8am} and ɑ ııꝰ {am}. Cappelli's work has given us reason to believe that, if medieval Latin is the substrate, then ɑ{a} is probably A, ııꝰ {m} is probably M, and ꝺ {8} has a good chance of being I. If that is so, then ꝺɑ ııꝰ {8am} maps nicely to the common Latin word IAM (which translates to the English "now," "already" or "besides").

The string ɑ ııꝰ {am} is more problematical. The word "am" exists in Latin corpora, but it is rare. We have to wonder whether this string is an abbreviation for a Latin word beginning with A and ending with M. If it is, a possible candidate is AUTEM (in English, "but," "however" or "also"), which occurs in Cappelli's work as and as . Alternatively, it could be the Voynich scribes' rendition of the Latin abbreviation ("amen" in Cappelli), which we would expect to occur hundreds of times if it turns out that the text of the Voynich manuscript is some form of scripture or liturgy.

The third most common "word," oꝛ {oe}, occurs 710 times. Here we have reason to think that o {o} is "o," but Cappelli gives us no clue as to what ꝛ {e} might be. In Latin there are relatively few two-letter words beginning with O. A possibility is "ŌS" ("speech," "mouth" or "face"), which would yield S, the fifth most common Latin letter, as the crib for ꝛ {e}.

Alternatively, oꝛ {oe} might not be a two-letter word at all but could be the Voynich scribes' version of one or more of the abbreviations , , or (all standing for "oz." or "ounce"). This would lead us to the ideas, firstly

that the manuscript is in some degree a work of chemistry or medicine; and secondly that the substrate languages might include Italian, since "ounce" is UNCIA in Latin but ONCIA, with plural forms ONCE or ONZE, in medieval Italian.

Likewise, ɑɋ {ae} could be the scribes' rendition of the Latin abbreviation , standing for AQUA ("water"), which Cappelli specifies as a pharmaceutical term; ɑƨ {ay} could be a form of , which stands for APUD ("at," "by" or "near"); and oƨ {oy} could be a variation of , short for OPUS ("work").

In similar fashion, we developed provisional mappings of the top twelve Voynich "words," either to Latin words listed in our reference corpus, or to Latin abbreviations which were listed by Cappelli. In some cases, we drew inspiration from our simplified v104 alphabet, which allowed us to ignore certain "marker" glyphs. The results are summarized in Table 11.2.

Table 11.2. Voynich manuscript: the twelve most common "words", and possible mappings to medieval Latin on the basis of the "Cappelli approach"

glyphs	v101	v104	v104 frequency	possible mapping	LatinISE corpus: frequency	English equivalent
ϑɑııƱ	8am	8aîΩ	0.4%	IAM	0.2%	now, already
ɑııƱ	am	aîΩ	0.4%	*ĀM. (AUTEM)*	*0.7%*	but, also
oɋ	oe	oe	0.4%	OS	0.0%	speech, mouth
ɑƨ	ay	aΨ	0.3%	A	0.9%	from, since
				AB	0.6%	from, since
oƨ	oy	oΨ	0.3%	Ō	0.1%	o, oh
				OP. (OPUS)	*0.1%*	work, job
ƨ	s	s	0.3%	ET	6.6%	and
ƈtoɋ	1oe	1oe	0.2%	EOS	0.2%	them
ƈtc9	1c9	1cΦ	0.2%	ATQUE	0.2%	and, and also
ɑɋ	ae	ae	0.2%	*AQ. (AQUA)*	*0.0%*	water
ƈtcϑ9	1c89	1c8Φ	0.2%	ATQUE	0.2%	and, and also

9	9	Φ	0.2%	CUM	1.6%	with
89	89	8Φ	0.2%	IUS	0.0%	law, right

Source: author's analysis. Mappings in normal font are whole words; italics are abbreviations. Latin corpus counts are from the medieval subcorpus of the LatinISE corpus, and refer to the numbers of words corresponding to the proposed Latin mapping (or in the case of abbreviations, to the expanded Latin mapping).

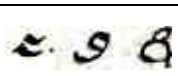
All of the top twelve Voynich "words" yielded medieval Latin words or abbreviations which were sufficiently common to be, in our judgment, plausible counterparts. The interested reader could certainly take this exercise further, and with a sufficient number of common Voynich "words," could probably construct cribs for at least the twenty most common glyphs, and thereby for at least 95 percent of the text in the Voynich manuscript.

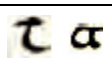
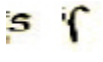
LATIN TO VOYNICH

To test this hypothesis from the other direction, we examined the Latin alphabet, starting with the most-common letters (E, I, U and so on) and continuing down the frequency table. For each letter, we looked in Cappelli's work for examples of Latin handwriting in which either the letter was written in full or was implicit in an abbreviation. Then for each letter, we searched for the Voynich glyph or glyphs which might reasonably be proposed to represent the letter or the abbreviation.

Table 11.3. The seven most frequent letters of possible counterparts among the Voynich glyphs

Latin letter	occurrences in top 1,000 words of LatinISE medieval subcorpus		selected examples of lower-case handwriting	expected v104 count	candidate Voynich glyphs and actual count			
	position	count			glyph	position	v104 key	v104 count
all	all	5,441,442		*168,262*				168,262

E	part of -QUE	20,605		*637*	ϙ	final	Φ	15,721
	part of -TER	12,510		*387*	ϼ	final	Ψ	6,173
	other internal	291,524		*9,015*	c	internal	c	13,969
	initial	225,989		*6,988*	ɑ c ϐ	initial	a c 8	6,759
	final	135,211		*4,181*	c cc	final	c ĉ	138
I	part of -IUS	20,042		*620*	ϙ	final	Φ	15,721
	part of -IS	44,345		*1,371*	ϙ	final	Φ	15,721
	other internal	325,690		*10,071*	ι ιι	internal	i î	8,319
	initial	130,602		*4,039*	ι ιι	initial	i î	185
	final	83,777		*2,591*	ι	final	i	7
U	part of QU-	121,603		*3,760*	ϙ	initial	Φ	1,908
	part of CUM-	21,673		*670*	ϙ	initial	Φ	1,908
	part of -US	66,772		*2,065*	ϙ	final	Φ	15,721
	part of -UR	11,163		*345*	ϑ	final	Ω	6,115
	other internal	248,043		*7,670*	ιι	internal	î	4,446
	initial	56,276		*1,740*	ιι	initial	î	4
	final	4,643		*144*	ιι	final	î	1

T	part of -TUM	12,908		*399*	⁀	final	îΩ	4,246
	other internal	172,406		*5,331*	ɑɑ	internal	1	5,893
	initial	40,951		*8,819*	ɑɑ	initial	1	9,713
	final	239,180		*7,396*	ɑɑ	final	1	68
S	internal	162,178		*5,015*	ȣ	internal	8	8,668
	initial	135,666		*4,195*	ȣ	initial	8	3,755
	final	151,278		*4,678*	ȣ	final	8	668
A	internal	205,180		*6,345*	ɑ	internal	a	11,717
	initial	99,521		*3,077*	ɑ	initial	a	2,714
	final	85,118		*2,632*	ɑ	final	a	55
N	part of CON-	5,474		*169*	ꝯ	initial	Φ	1,908
	part of -NUS	6,774		*209*	ꝯ	final	Φ	15,721
	other internal	173,968		*5,379*	⁀	internal	î	4,580
	initial	81,037		*2,506*	⁀	initial	iΩ	0
	final	90,435		*2,796*	⁀	final	iΩ	1,740

Source: author's analysis; Latin handwritten letters and abbreviations from Cappelli (1929).

actual glyph count substantially in excess of expected count

actual glyph count substantially below expected count

Our criteria were at least threefold:

- the Voynich glyph should preferably have some morphological resemblance to at least one handwritten variant of the Latin letter or abbreviation

- the frequency of the Voynich glyph should approximate, within reasonable bounds, to the frequency of the Latin letter or abbreviation
- the frequency of the Voynich glyph in each position relative to the word (initial, internal and final) should approximate to the frequencies of the Latin letters in the respective positions. For example, if the letter E occurs most often in an internal position, any Voynich glyph that we propose to represent E must also occur with an acceptable frequency in an internal position (and as we shall see in a moment, this was the toughest criterion to meet)

We conducted this analysis for the nineteen most common letters in our reference corpus of medieval Latin (which together accounted for 98.9 percent of all the letters in the corpus). The provisional results of this analysis, for the top six letters—E, I, T, U, A, and S—are set out in table 11.3.

We can see from Table 11.3 that for the top six Latin letters, we could generally find an arguable visual counterpart or counterparts among the Voynich glyphs. Moreover, if the shape of the letter differed between the initial, internal and final positions, or if the letter was concealed in one or more abbreviations, there were usually several glyphs that would approximate to the variations in the shape, or to the abbreviations

However, the frequencies of the Voynich glyphs were often out of line with what we would expect in a document derived from a text or texts in medieval Latin.

Some examples may illustrate our dilemma.

THE LETTER “E”

By our calculations, in medieval Latin texts the most common letter was E. In our reference corpus, it accounted for 12.6 percent of all the letters.[6] Two of the three most common words in our corpus were ET (“and”) and EST (“he, she or it is”). In a Latin text of a similar length to the Voynich

manuscript, we should expect the word ET to occur at least 2,500 times. If the word was transcribed in full, we should be looking for a two-glyph "word" with this frequency. We did not find such a "word"; but the Voynich scribes may have abbreviated it to a single glyph, for which we have already nominated {s} as a candidate; and it may not be the only one.

From the movie *Gladiator*, we recall the chilling address of the Spanish gladiator and former general Maximus, in the Roman arena, to the emperor Commodus:

> My name is Maximus Decimus Meridius, commander of the Armies of the North, General of the Felix Legions and loyal servant to the true emperor, Marcus Aurelius. Father to a murdered son, husband to a murdered wife. And I will have my vengeance, in this life or the next.[7]

If the screenwriters had been sufficiently imaginative, they might have allowed Maximus to speak Latin, the only language in which it would be permitted to address the emperor, and he would have said, perhaps with the accent of his native province of Extremadura:

> Mihi nomen est Maximus Decimus Meridius, praefectus copiarum septentrionis, dux legionum Felix, servus fidelis veri imperatoris Marci Aurelii. Pater filii interfecti, vir mulieris interfectae. Atque eorum mortem ulciscar, in hac vita aut in postera vita.[8]

In this short speech, which uses 220 letters in written Latin, Maximus would have used the letter E twenty-six times.

Likewise, if we could map the approximately 160,000 glyphs of the Voynich manuscript into Latin on a one-to-one basis, we should expect the letter E to occur about 20,000 times. Not only that: it should occur about 9,000 times as an internal glyph, 6,500 times in the initial position, and 4,500 times in the final position.

It is not straightforward to identify a counterpart glyph in the Voynich manuscript. There is no Voynich glyph which is both simple to write and

has the distinctive closed loop of a handwritten E, at least not in the right place. The common glyphs that come close to resembling an E are {c} and {a}.

In gross terms, we have a sufficient number of these two glyphs. In the whole manuscript, together they occur 28,248 times. But they are not in the right places. Of the total, 25,238 are in internal positions which is ample for internal E's; but there are far too few in initial and final positions. To represent the initial E's that we should expect in a normal Latin text, we would need to enlist the glyphs {9} and {8}—which do indeed resemble some variants of the handwritten initial "e," but we have no idea how rare or common those variants were. To represent the final E's, we have no obvious candidate, except for the relatively few instances in which the E was part of the suffix -QUE, in which case the glyph {9} would have been the logical choice.

In short, for the most common letter in the Latin alphabet it is already difficult to imagine how the Voynich scribes would have handled the transcription. If we hold onto our hypothesis that they were working from Latin texts, we must conclude that they were free to select multiple glyphs to represent the letter E (and other letters, as we shall see in a moment). Furthermore, they could use one glyph for an initial E and another for a final E. One also suspects that they could simply omit both initial and final E's. If so, they were, or they were instructed to be, radical abbreviators; and yet they were not short of parchment.

THE LETTER "I"

The letter I occurred in our reference corpus with a frequency of 11.1 percent. In a medieval Latin text, we should expect to find this letter often in the initial position, in words such as IN (English "in") and ID ("it"); in the final position, in words like QUI ("how" or "why") and SI ("if"); and internally in words like QUIA ("because").

The short speech of Maximus the gladiator, had he spoken in Latin, would have used the letter I thirty-three times. Likewise, if we are to be at all successful in our mapping of the Voynich manuscript into Latin, we need

the glyph or glyphs representing the letter I to occur about 17,500 times: of which 11,000 should be internal, 4,000 initial and 2,500 final.

The Voynich glyph which most resembles the letter I is ı, which is even assigned the letter {i} in the v101 alphabet. But in the v101 transliteration, ı occurs only 264 times; and in what looks like a doubled form ıı {I}, only ninety-eight times. So, to assemble enough I's for an intelligible Latin text, we have to look among the other Voynich glyphs. Here are some possibilities:

- The glyph o {9}, in the final position, occurs 15,020 times in v101. Cappelli tells us the Latin suffix 9 can represent the suffix -IS, which yields an internal I. We expect about 1,300 instances of this suffix, so we have enough glyphs for this purpose.
- The glyph ıi {n} occurs 1,739 times, nearly always in the final position In our v104 alphabet we broke up this glyph into an ı [i] followed by a i {N}.
- The glyph ƒ {p} occurs 989 times, again predominantly as a final glyph. In v104 we broke it into an ı {i} and a f [*].

For the internal I's, the only device which could yield a sufficient number was to liberate the instances of ı {i} and ıı {I} that were imprisoned in glyphs such as ııi {m}, ƒ {p} and their variants. That would yield in the order of 10,000 I's; but we would have to abandon the idea that ııi {m}, ƒ {p} and their variants are single glyphs. For initial and final I's, the Voynich glyphs offer no candidate with the right number of occurrences. Again, in our scenario of transcription from Latin source documents, it is as if the Voynich scribes simply omitted the Latin I's in these positions.

THE LETTER "U"

The U was the third most common letter in our reference corpus, with a frequency of 9.7 percent. If we want a plausible Latin transliteration of the Voynich manuscript, we are looking for about 15,000 glyphs representing U: of which about 13,500 should be internal and about 1,500 initials. A typical Latin text of this length should contain hundreds of instances of words such as QUOD ("which"), CUM ("with"), QUI ("how," "why" or "that which"), QUAE (feminine "qui") and QUAM ("how much").

Our search for the U among the Voynich glyphs involved some excavation. We suspected that the Latin U was buried within some of the glyphs, either as a shape resembling a U, or as an implied U within an abbreviated prefix or suffix. Our candidates included the following:

- Perhaps the prime candidate is the glyph {1} in the initial position (9,713 instances in v104); this glyph does not visually resemble a U but has the binary quality that makes us suspect it to represent the Latin QU.
- The glyph {9} in the final position (15,397 instances in v104, of which about 4,000 should be sufficient to represent the suffixes -US, -IUS, -MUM, -MUS, -NUS, -QUE, or -TUM).
- The glyph {9} in the initial position (1,674 instances in v104, of which about 600 would be all we need to represent the expected prevalence of the prefixes CUM- or CUN-)
- The glyph {m} in the final position (in v104 we have broken up this glyph to extract the imprisoned , thinking that it may represent a U).

Even with all these candidates, we will still be short of the number of U's that we expect in a Latin document. We would have expected a Latin text of this length to contain at least 600 instances of the word UT ("in order to"); but we have practically no Voynich glyphs in the initial position that look like a Latin U.

THE LETTER "T"

By our calculations, the letter T was the most common consonant in medieval Latin (as it is in modern English and other modern European languages). In our reference corpus, T accounted for 8.6 percent of all the letters. In a Latin text, we should expect to find the T in ubiquitous words such as ET, EST, and UT. Maximus, to make the emperor tremble, would have used the T sixteen times. We need about 13,500 instances of glyphs representing T: about 5,500 internal, 1,000 initial and 7,000 final.

Here there is one Voynich glyph which has the right shape and the right frequency. It is the glyph {1}, which occurs 15,655 times in our v104 transliteration (in which we combined it with {2} and its variants). To the possible objection that it does not resemble a T, here is one of Cappelli's examples of a handwritten lowercase Latin T . To the objection that we have already nominated it to be a possible QU, we could respond that we have enough instances of it to be both T and QU.

In case we need more T's, we could look for glyphs resembling the Latin (the suffix -TIS) or (the suffix -TER). Here the candidates could include the glyph {y} with 6,635 instances in v101, {z} with 505 instances and {p} with 984 instances.

The main anomaly in our proposed mapping of T is that we do not have enough candidate glyphs in the final position. Even all the variants of {p} will only yield about 1,000 of the 7,000 that we expect. Increasingly, it looks as if the Voynich scribes (if they were working from documents in Latin) were free to replace a final Latin letter with any one of several glyphs of their choice, but with an overwhelming preference for the final {9}.

THE LETTER "S"

The letter S had a frequency of 8.3 percent in our reference corpus. The "s" should appear in common Latin words such as EST ("is"), SED ("but"), or SI ("if"). Our target is to find about 12,000 instances of glyphs that could represent a Latin S.

We have one strong candidate that stands on its own: the glyph 8 {8}, in which it does not take much imagination to see the shape of an S, and which occurs 13,208 times in the v104 transliteration.

THE LETTER "A"

With one more example, we might reach a reasonable level of confidence that this approach could bear fruit. The letter A accounts for 8.9 percent of all the letters in the medieval Latin corpus. The speech of Maximus to the emperor Commodus would have had fifteen instances of A. In our mapping of the Voynich manuscript, we should be looking for about 14,000 A's.

Here we have a glyph which looks about right: a {a}. There are 14,497 instances of this glyph in both the v101 and the v104 transliterations.

OTHER LETTERS

We continued this approach for the next thirteen most frequent letters in the Latin alphabet. In each case there seemed to be a reasonable possibility of a mapping from a glyph or several glyphs in the Voynich manuscript.

We have to admit that this approach does not provide a strong basis for mapping the Voynich "gallows" glyphs, which, with the exception of f {h} and p {k}, are relatively rare. These glyphs do not greatly resemble any Latin letters or abbreviations as documented by Cappelli. There are several possibilities which are open to testing, including the following:

- The "gallows" glyphs could represent doubled Latin letters, in which case f {h} and p {k} could stand in some order, or interchangeably, for "ll" and "ss," which are the two most common doubled letters in Latin.[9]
- The "gallows" glyphs could represent combinations of Latin consonants, in which case f {h} could stand for NT and p {k}

could stand for ST, which are by far the most-common pairs of consonants in Latin.[10]

- The “pedestal” glyphs seem to be “plain gallows” glyphs sitting on a stretched version of the glyph , which we have suspected as mapping to the Latin QU and/or T. In that case (since combinations of three consonants are extremely rare in Latin) these glyphs could be some combination of the letters T or QU plus an implicit vowel and a final consonant.
- The “grand piano” glyphs, which are “plain gallows” glyphs sitting on the unassigned glyph , could be some combination of the letters T or QU plus an implicit vowel, another consonant, and the letter E or C.

MAPPINGS

In any case, at this point it was time to put this approach to the test, with some mappings. Up to this point in the exercise, we had kept in mind the differences between the glyph frequencies in Currier’s languages A and B; but none was so great as to suggest two separate mappings between Latin letters and Voynich glyphs. However, for a good test of the approach, it seemed advisable to select passages from both languages A and B. For this purpose, the first three lines of page f001r (the opening page of the manuscript) seemed reasonable representatives of Language A, and the first three lines of page f026r served as a sample of Language B.

Broadly speaking, our approach in these mappings was as follows:

- Let any initial {9}, {s}, or {y) map to one of the prefixes CO-, COM-, or CON-, whichever fitted better
- Let any final {9} map to the suffix -US
- Let any final {89} map to the suffix -IUS
- Let any stand-alone {9} map to the word CUM
- Let any final {e} map to the letter -S

- Let any final ʔ {y} map to the suffix -RE; in an internal position, let this glyph map to a P
- Let ℍ {h} and 𝕂 {k} map to M
- Let all the other "gallows" glyphs map to a catch-all "*," indicating a gap in our mapping
- Let all other glyphs map to the letter most closely resembling the glyph

The first, and of course provisional, results are shown in Table 11.4.

It will be apparent that these transliterations are not Latin words as we know them. Nor, on a first attempt, would we expect them to be. There is a proliferation of Latin suffixes, and to a lesser extent, of Latin prefixes, but that follows from the methodology. The words are not those that we remember from our Latin textbooks. The interested reader may be tempted to experiment with permutations of the mapping.

Table 11.4. The Voynich manuscript: a test of the "Cappelli approach" - a semi-heuristic mapping from Language A and Language B to Latin

page f001r, first line (Currier language A)

Voynich v101 Unicode	ℍacɾoʔ oℍaɾ aʔ olℍaıı̇ ch͡oɾ ch͡oʔo Ξʔch͡ʔ o ℍoʔ ch͡oɾ8o
v101	fa19s 9hae ay oham 2oe 2oy9 Kscs 9 hoy 2oe89
v104	fa1Φs Φhae aΨ ohaîΩ 1'oe 1'oy9 h1scs Φ hoΨ 1'oe8Φ
suspected Latin abbreviations	fa19s Φhae aΨ ohaîΩ 1'oe 1'oy9 h1scs Φ hoΨ 1'oe8Φ
with abbreviations expanded	fa1Φs CONhae aΨ ohaîΩ QU'oe QU'oΨUS h1scs CUM hoΨ QU'oe8US
other letters mapped according to vowel probability and frequency	
Latin, classical	VUEAP CONTUR UN ITUCO QU'IR QU'INUS TEPMP CUM TIN QU'IRSUS

Latin, 14th century	GUEAP CONTUN UR ITUCO QU'IN QU'IRUS TEPMP CUM TIR QU'INSUS

page f026r, first line (Currier language B)

Voynich v101 Unicode	ﻩﻩﻩﻩﻩﻩﻩﻩﻩ
v101	g2coh9 o8aZ 4o9 <> of2o8 19g1c9 9g1c89 an <> 1ofo1g89
v104	g1'cohΦ o8aiΨ ΔΦ <> of1'o8 1Φg1cΦ Φg1c8Φ aiΩ <> 1ofo1g8Φ
suspected Latin abbreviations	g1'cohΦ o8aiΨ ΔΦ <> of1'o8 1Φg1cΦ Φg1c8Φ aiΩ <> 1ofo1g8Φ
with abbreviations expanded	g1'cohUS o8aiΨ ΔUS <> of1'o8 QUΦg1cUS CONg1c8US aiΩ <> QUofo1g8US
other letters mapped according to vowel probability and frequency	
Latin, classical	GM'EISUS ITAiN LUS <> IFM'IT QUOGMEUS CONGMETUS APU <> QUIFIMGTUS
Latin, 14th century	VM'EISUS ITAiR DUS <> IHM'IT QUOVMEUS CONVMETUS APU <> QUIHIMVTUS

Sources: author's analysis. Image: *Beinecke Rare Book and Manuscript Library*. Notes: <> denotes break in text where an illustration intervened. Frequencies in classical Latin by courtesy of Stefan Trost Media. Frequencies in 14th century Latin calculated by author from Dante Alighieri's

CANNABIS

We then tried a test of this approach on a "herbal" page where we had a possible identification of the plant, namely page f016r. In Chapter 8 we suspected that the illustration on this page might represent the *Cannabis* genus, commonly known as hemp. The word CANNABIS is Latin, and occurs in the medieval LatinISE subcorpus; so, if page f016r was derived from a Latin document, we might hope for a word representing CANNABIS to appear somewhere.

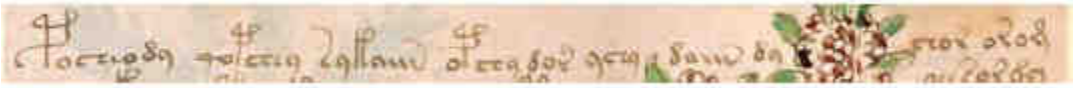
We should already know what we are looking for. The final IS might be represented by a terminal ﻩ {9); the second and fourth (or fifth) glyphs

should be ɑ {a). The N might or might not have been doubled in the original. The B is approximately the fifteenth most common internal letter in medieval Latin.

The nearest possibility is, as in Chapter 9, the first glyph string ⁋oc꜀co89{go1co89}, but then the glyph ⁋ {g} improbably has to represent C and o {o} has to represent A. This does not necessarily invalidate the ideas behind the mapping; it may be simply that the plant is not cannabis, or that the text has nothing to do with the illustration.

Table 11.5. The Voynich manuscript: another test of the "Cappelli approach" - a semi-heuristic mapping from the "cannabis" page to Latin

Folio f016r, first line (Currier language A)



Voynich v101 Unicode	⁋oc꜀co89 ɬo⁋c꜀c9 2ƍ⁋ɑıı∀ o⁋c꜀9 8o? 9c꜀9 8ɑıı∀ 89 <> c꜀o? o?oɣ
v101	go1co89 4oj1c9 s9ham og19 8oy 919 8am 89 <> 1oy oyop
v104	go1co8Φ Δg1cΦ sΦhaîΩ og1Φ 8oy Φ1Φ 8aîΩ 8Φ <> 1oΨ oΨoΠ
suspected Latin abbreviations	go1co8Φ Δg1cΦ sΦhaîΩ og1Φ 8oΨ Φ1Φ 8aîΩ 8Φ <> 1oΨ oΨoΠ
with abbreviations expanded	go1co8US Δg1cUS sΦhaîΩ og1US 8oTER Φ1US 8aîΩ 8US <> 1oTER oΨoT
other letters mapped according to vowel probability and frequency	
Latin, classical	GIMEITUS LGMEUS QUOSAPU IGMUS TITER CONMUS TAPU TUS <> QUITER INIT
Latin, 14th century	VIMEITUS DVMEUS BOSAPU IVMUS TITER CONMUS TAPU TUS <> QUITER IRIT

Sources: author's analysis. Image: *Beinecke Rare Book and Manuscript Library*. Notes: <> denotes break in text where an illustration intervened. Frequencies in classical Latin by courtesy of Stefan Trost Media. Frequencies in 14th century Latin calculated by author from Dante Alighieri's *Monarchia, I-III*.

TAURUS

We made a third test of the "Cappelli approach" with the three labels that, in Chapter 10, we thought might be associated with the Pleiades star group in the constellation of Taurus. Here we were looking for a medieval name for the Pleiades, or for any of the seven stars within the Pleiades, or for the bright star Aldebaran nearby.

Under the Cappelli approach, the mappings of the three labels were as follows:

- For the label next to the star group, ꞔoaʔ9 {8oay9}, possible mappings included SOAPUS, SOATERUS, or SOAREUS, assuming that the initial ꞔ{8} had to be an S, and depending on how we interpreted the penultimate ʔ {y}. We might see therein the Romanian word SOARE ("sun"), or that might be a coincidence.
- The label for the bright star nearby, ꞔcoɋꞔa9 {81oe8a9}, mapped to STOS SAUS, STOS SAOS, or STOS SAIS.
- The label for the connecting line to the sun, oaɋccoɋ {oae1coe}, which we suspected to be two words, yielded the possible mappings OAS TCOS or OAS QUCOS.

None of the mappings could be found as a word or words in the LatinISE corpus, and none of them evoked any known name associated with the Pleiades, with the individual stars of the cluster, or with the star Aldebaran.

NEXT STEPS

The Cappelli approach offers some promise. We have proposed that many of the mappings from Voynich glyphs to Latin letters are not unique, and therefore they permit variants that can only be tested by trial and error. Perhaps a reader of this book, armed with the "Cappelli approach" but better versed in medieval Latin than this author, will find narrative meaning in these or other passages of the Voynich manuscript.

CHAPTER 12

CANTIGAS

In Chapter 5, we alluded to Capt. Currier's "functionality of the line" and we drew the inference that the Voynich manuscript was probably not a work of prose, but more probably a work of poetry.

We can test this hypothesis with a bit of mathematics.

Suppose that we have a document written in prose, such that the text wraps from each line to the next and there is no hyphenation of words. Let us say that the document is in medieval Latin. Then, for any given letter:

- The probability that the line ends in that letter is simply the frequency of the letter in the final position in the language corpus.[1] For example, the probability that a line ends in "t" is the frequency of the final "t" in medieval Latin, which is 18.4 percent.[2]
- The probability that the next line ends in that letter is the same: in the case of "t," 18.4 percent.

- The probability that two consecutive lines end in that letter is the square of the frequency: in the case of "t," 0.184 × 0.184 = 3.4 percent.

The probability that two consecutive lines end in the same letter (no matter what that letter may be) is the sum of these probabilities for all the letters in the alphabet: which, for letters in medieval Latin, is 10.8 percent.

In other words, if we may use the term "rhyme" to apply to the recurrence of the final letter: two consecutive lines of medieval Latin prose will unintentionally rhyme with a probability of 10.8 percent.

On this definition and by our calculations, the probabilities of rhyming consecutive lines of prose in selected medieval languages other than Latin are as follows:

Middle English	6.4 percent (based on the Auchinleck Manuscript)
Middle French	7.5 percent (based on *La Farce de Maistre Pathelin*)
medieval Italian	16.9 percent (based on the OVI corpus)
Old Albanian	18.4 percent (based on Gjon Buzuku's *Meshari*).

The profusion of rhyming lines in Old Albanian follows from its vocabulary. In *Meshari*, four of the eleven most common words have the same ending: "teh," "meh," "seh," and "ʃeh." If these words fall by chance at the endings of two consecutive lines, the lines will seem to rhyme even if the rhyming was unintentional.

Not knowing the underlying language or languages of the Voynich manuscript (if such exist), we could expect that if the manuscript is a work of prose in a medieval European language, with text wrapping and no hyphenation, then two consecutive lines should "rhyme" (that is, end in the same glyph) with a probability in the range between those of Middle English and Old Albanian.

In the v101 transliteration, there are 5,367 lines of text in the Voynich manuscript (using the term "line" to include horizontal lines, circular lines, radial lines, and standalone labels). Each of the lines is identified by its folio number and its position on the page (in some cases, such as circular lines, starting from an arbitrary point on the page). For example, line 8 on folio f016r is identified as line f016r.08. Accordingly, we can determine, with a small degree of arbitrariness, which line follows which.

If the 5,367 lines of the Voynich manuscript are prose, we should expect between 10 and 18 percent of the lines to "rhyme": that is, between about 550 and about 1,000 lines.

Now we may apply the Excel COUNTIF function to the Voynich manuscript and see how many lines actually rhyme. The results are shown in Table 12.1.

In the transliteration to the v101 alphabet, we see that there are 1,199 lines which rhyme with the preceding line: more than the largest number that we would have expected from a medieval European language. Of the corpora that we studied, the Voynich manuscript is most similar in this respect to *Meshari* in Old Albanian—which is a translation of the Catholic liturgy. Furthermore, 1,224 lines rhyme with the second preceding line; we may recall that this is the rhyming structure of Dante Alighieri's *La Divina Commedia*. In total (and excluding double-counting), there are 2,017 lines of the Voynich manuscript which rhyme with either the preceding line or with the second preceding line. This is equivalent to over 37 percent of the manuscript.

Table 12.1. Voynich manuscript: numbers of lines which rhyme

Currier language	total	rhyming with preceding line		rhyming with 2nd preceding line		rhyming with both		total rhyming	
v101 transliteration									
A	1,818	336	18.5%	350	19.3%	97	5.3%	589	32.4%
B	3,549	863	24.3%	874	24.6%	309	8.7%	1,428	40.2%
all	5,367	1,199	22.3%	1,224	22.8%	406	7.6%	2,017	37.6%
v104 transliteration									
A	1,818	336	18.5%	350	19.3%	97	5.3%	589	32.4%

B	3,549	936	26.4%	943	26.6%	351	9.9%	1528	43.1%
all	5,367	1,272	23.7%	1,293	24.1%	448	8.3%	2,117	39.4%

Source: analysis by author. Pages in Languages A and B as identified by Currier (1976) and Davis (2022)

The v101 alphabet gives separate keyboard assignments to minor variations in the glyphs: for example, the common final glyphs ⟲ {m} and ⟳ {n} differ by one quill stroke. In the simplified v104 alphabet that we have proposed, variations such as these are subsumed within a single glyph, which means that more lines will rhyme. This is indeed the case: as Table 12.1 illustrates, in the v104 transliteration we have 2,117 lines, or almost 40 percent of the manuscript, with a rhyming structure.

IF IT IS POETRY …

If the Voynich manuscript is in whole or in substantial part a work of poetry, there are a number of implications for our research.

Firstly, we can afford to think less about the letter frequencies of glyphs and letters than we have done so far. As we saw in Chapter 5, in Dante's *La Divina Commedia* all of the thirteen most common letters of medieval Italian have substantial divergences in frequency from the corpus of Italian language of that era. These divergences occur in particular when the letter is at the beginning or end of a line. This is as we could expect, given that a poet is subject to the conventions of poetry.

Secondly, we would not worry if the Voynich glyphs do not appear to include representations of numbers. Poetry normally does not use numerics; if there are numbers in the text, they will be expressed in words.

Thirdly, we might expect a substantial prevalence of repeated "words" (strictly speaking, glyph strings). Repetition is a conventional tool of poetry, for dramatic impact and for preservation of the rhythm. Shakespeare's *Macbeth* provides a good example:

> Tomorrow, and tomorrow, and tomorrow,

> Creeps in this petty pace from day to day,
> To the last syllable of recorded time.[3]

Likewise, the Voynich manuscript has many sequences of repeated "words":

> While simple (i.e. double) reduplication occurs more than 250 times in the manuscript, triple (or more) reduplication occurs about 10 times.[4]

An example of repetition of the glyph string ﬓ ıı◌ {8am} is shown in Figure 12.1 below, from Voynich page f005v, line 5. There are at least seventeen instances of this particular repetition in the Voynich manuscript.

If, as we suggested in Chapter 11, this string might correspond to "iam" ("now" or "already") in Latin, it is in good company. In the LatinISE corpus, there are 166 instances of the phrase "iam iam" or "iam iamque," with the sense of an emphatic "now." In the year 1486 Augustin Tünger, a prosecutor at the Bishop's Court in Constance, a free city of the Holy Roman Empire, wrote in *Facetiae latinae et germanicae*:

> iam iam flammam extinctam esse
> ["now, now, the flame is to be extinguished"][5]

A possible example of three consecutive instances of a glyph string is shown in Figure 12.2, in which the glyph string ◌̂ch◌oq {2coe} appears to occur three times:

In any case, for any given language that we may propose to underly the Voynich manuscript, we are not greatly constrained by the frequencies of letters in that language. Even if we cannot make a sensible mapping on the basis of glyph and letter frequencies, it does not mean that the language is ruled out. The Voynich manuscript may be written in the language in question, but in a poetic form which transcends the conventional rules of prose.

For example, in Chapter 5 we had high hopes for medieval Italian as Currier's Language B, on the basis of the correlation between the glyph and letter frequencies. We tried a transliteration of the first few lines of page f026r into medieval Italian, using the "Gold Bug" approach; but it did not produce sensible results. However, we were using the OVI online corpus of medieval Italian, which we suppose is based on administrative and ecclesiastical texts. Language B could still be medieval Italian, but perhaps it resembles the Italian of Dante Alighieri rather than that of the online corpora as used by scholars today.

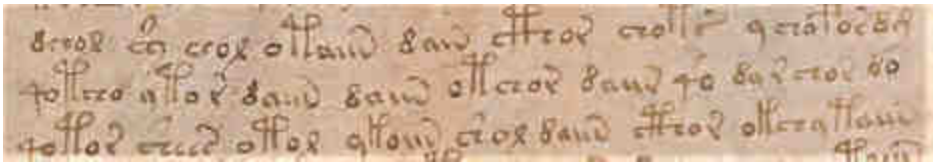


Figure 12.1. Voynich manuscript, extract from page f005v, lines 2-4, showing repetition of the glyph string {8am}. *Beinecke Rare Book and Manuscript Library*

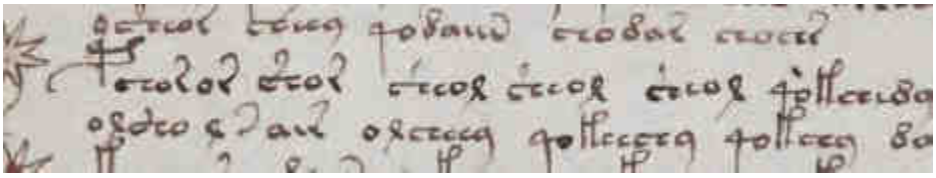


Figure 12.2. Voynich manuscript, extract from page f104v, lines 26–28, showing possible tripling of glyph string {2coe}. *Beinecke Rare Book and Manuscript Library*

Fourthly, notwithstanding our hereto acceptance of Currier's languages A and B, we cannot exclude that a poetic structure could give the statistical appearance of multiple languages, but in reality, could conjure from a single language a multiplicity of rhythms and sonorities. If, not knowing the English language or the Latin script, we read the following verse:

> Corrina, Corrina, gal you been on my mind
> Corrina, Corrina, gal you been on my mind
> When I think about you babe
> I just can't keep from cryin'[6]

and then this one:

> And the all-night girls, they whisper escapades out on the D train You can hear the night watchman, click his flashlight, wonder if it's really him or them that's insane,[7]

would we imagine that they were both written in the same language, by Bob Dylan?

CANTIGAS

We are impelled therefore to search for a corpus of medieval poetry, as contemporary with the Voynich parchment as we can find, and preferably in a Romance language which will fall within our first or second circles of target languages. If it is the wrong language, a few attempts at mapping will possibly point us in the right direction. We propose therefore to devote this chapter to an examination of the *Cantigas d'Amigo*.

The *Cantigas d'Amigo* and its companion *Cantigas d'Amor* were love songs, written in Galician-Portuguese during the thirteenth and fourteenth centuries (and therefore predating the Voynich parchment by up to a hundred years). We have to hope that the difference in time does not exclude the *Cantigas* as an inspiration or a precursor of the Voynich manuscript; in the 2020s, if we are lucky, we might still come across snatches of lyrics or melody from the roaring 1920s.

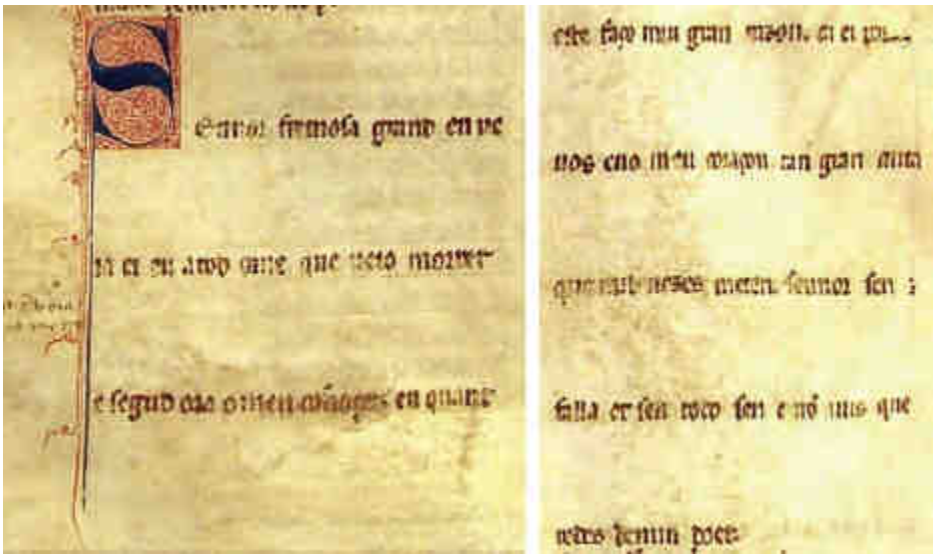


Figure 12.3. “Senhor fremosa, grand’enveja hei,” by Vasco Praga de Sandim, first verse, from *Cancioneiro da Ajuda. Palácio da Ajuda*

All of the *Cantigas* were written and performed by male, poets and singers. The *Cantigas d’Amor* were performed by the singer speaking in the first person; the *Cantigos d’Amigo* were performed as if voiced by a woman.[8]

Figure 12.3 shows the first verse of the *Cantiga de Amor* “Senhor fremosa, grand’enveja hei,” by Vasco Praga de Sandim, providing an example of the ornate first letter of the cantiga, and of the spaces left for the insertion of musical notation (which in this instance, were never added).

For our understanding of the structure, style, and vocabulary of the *Cantigas*, we are indebted to Rip Cohen and in particular to his magisterial critical edition and English translation of the *Cantigas d’Amigo*.

Here is an example of a short *Cantiga d’Amigo*, consisting of eleven lines in four verses, written and performed by Paio Soarez de Taveirós, probably in the early thirteenth century, with this author’s translation to modern English:

O meu amigo, que mi dizia	The friend of mine, who used to say
que nunca mais migo viveria,	that never more he’d stay with me,

par Deus, donas, aqui é ja	by God, ladies, he's here again.
Que muito m'el avia jurado	How much to me he had declared
que me non visse mais, "a Deus grado,"	he'd never see me more (thank God),
par Deus, do<nas, aqui é ja>	by God, ladies, he's here again.
O que jurava que me non visse,	The one who swore he'd see me not,
por non seer todo quant'el disse,	for nothing was as it should be,
par Deus, donas, aqui <é ja>	by God, ladies, he's here again.
Melhor o fezo ca o non disse:	Better his deed than what he said,
par Deus, <donas, aqui é ja>	by God, ladies, he's here again.

Table 12.2. *Cantigas d'Amigo*: the twenty-nine most common letters, compared with a medieval Galician-Portuguese corpus

letter	Corpus Xelmirez	Cantigas d'Amigo
E	5.2%	14.7%
A	5.2%	10.4%
O	5.1%	9.3%
R	5.1%	7.0%
S	5.1%	6.8%
N	4.9%	6.6%
I	4.8%	6.3%
U	4.7%	5.9%
M	4.6%	5.1%
D	4.8%	4.6%
L	4.4%	2.5%

Q	3.5%	2.4%
V	2.6%	2.4%
T	4.9%	2.3%
P	4.1%	2.1%
C	3.1%	1.8%
G	2.0%	1.7%
H	2.9%	1.3%
F	1.9%	1.2%
B	1.6%	0.8%
Z	1.3%	0.7%
J	0.9%	0.5%
Á	0.8%	0.3%
Ó	0.5%	0.2%
É	0.5%	0.2%
Ç	0.4%	0.2%
X	0.2%	0.1%
Õ	0.0%	0.1%
Í	0.1%	0.0%
others	38.8%	0.0%
Total	100%	100%

Source: analysis by author. Sources of data: *Cantigas* from Lopes, Graça Videira; Ferreira, Manuel Pedro et al. (2011-); *Corpus Xelmirez* from Instituto da Lingua Galega

frequency is over 2 percentage points above that in corpus

The rhyming structure, exemplified by "dizia" and "viveria," seems intrinsic to the *Cantigas d'Amigo*. To test this hypothesis, we extracted 1,321 lines of the *Cantigas*, by sixteen different poets, from Rip Cohen's *Critical Edition*. We imported these lines to an Excel database, which enabled counting of the rhyming instances. Among the 1,321 lines, there were 535 which rhymed (in the sense of having the last two letters in common) with either the preceding line or with the second preceding line. That represented a rhyming incidence of 40.5 percent of the sample: a very similar incidence to what we had seen in the Voynich manuscript.

The next step was to calculate the letter frequencies of the *Cantigas d'Amigo*, for comparison with the frequencies in medieval Galician-Portuguese as a whole. Our results are shown in Table 12.2.

Here we can see that the letter frequencies in *Cantigas d'Amigo* are in some respects substantially different from those in medieval Galician-Portuguese in general. Compared to the corpus, *Cantigas* are much more likely to use the common letters "e," "a," and "o"; this may be a consequence of the poetic and rhyming structure. The eleven lines of the cantiga "O meu amigo que mi dizia" by Paio Soarez de Taveirós, as we saw a few pages back, end in every case with "a," "e," or "o."

ABBREVIATIONS IN THE *CANTIGAS*

If we imagine a situation in which the scribes of the Voynich manuscript meticulously mapped a document or documents written in Galician-Portuguese into the glyphs that we see today, we have to consider that they would have had before them documents that were to some extent, abbreviated.

Just like the scribes of medieval Latin, the writers of the *Cantigas* were accustomed to economy in the use of parchment. They shared a convention of abbreviation that, to some degree, overlapped the conventions of Latin abbreviation to which we alluded in Chapter 11.

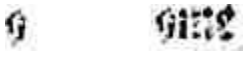
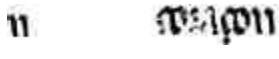
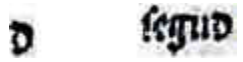
In Table 12.3 we have assembled a selection, by no means comprehensive, of abbreviations which may be found in just two of the *Cantigas*.

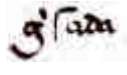
Some observations:

- The abbreviation sign ⁹ was used commonly at the end of a word, to signify "-os" or "-us" (as it was in medieval Latin). We found that a similar sign was used sometimes at the beginning of a word to signify "ho-" (for which we know of no counterpart in Latin).

- A superscript or was used at the ends of words to signify "-m," and occasionally within a word to signify "-n-."
- An "m" at the end of a word was often written "n," as in ("coraçom" or "heart")
- The syllable "que," whether a whole word or part of a word, was universally abbreviated to a single "q" with a superscript , very similar to that used elsewhere to mean "-m"
- A superscript resembling two dots or the German ümlaut could stand variously for "-er" or "-ra"

Table 12.3. Cantigas: selected abbreviations in the original manuscripts

meaning	position in word	symbol	examples
-er	final		dizer
-er	final		per
eu	final		hoj'eu
ho-	initial		home
-m	final		bem
-m	final		gram
-m	final		nom
-m	final		coraçom
-n-	internal		nunca
-ndo	final		segundo
-or	final		por
-or	final		por
-os	final		ambos

-ra-	internal			grande
-ue	final			que
-ue	internal			quer
-uel	final			aquel
-ui-	internal			guisada

Sources: Vasco Praga de Sandim: “Senhor fremosa, grand’enveja hei;” João Garcia de Guilhade: “A Dom Foam quer’eu gram mal,” retrieved on August 23, 2021 from Lopes, Graça Videira; Ferreira, Manuel Pedro et al. (2011-)

The abbreviation signs in the *Cantigas* evoke the Voynich glyph ⁹ {9}, which is ubiquitous at the beginnings and ends of glyph strings. We may also suspect that if the scribes of the Voynich manuscript were copying from documents in Galician-Portuguese, they may frequently have come across signs or superscripts with more than one meaning, or with a meaning dependent on context. If so, they may not have had the time (or may not have received instructions) to do anything other than transcribe the sign into a corresponding glyph. If that was the case, we must likewise be prepared to interpret the Voynich glyphs according to their position in the glyph string, or their context, or both.

MAPPING FROM VOYNICH

We now return to our search for a concordance with the Voynich glyphs, at least the most-common ones.

A prudent first step, as in our examinations of other languages, would be to identify the most-common words in *Cantigas d’Amigo*. If there is any hope of a concordance, some of those words should resemble some of the common Voynich “words,” in length, in character sequence, and in some approximation of frequency.

The twenty most common words in *Cantigas d'Amigo*, and the twenty most common "words" in the Voynich manuscript (with some variants) are summarized in Table 12.4.

The most striking possibility of a concordance lies in the two most common "words" in the Voynich manuscript: ƈɑ⁀ƞ {8am} and ɑ⁀ƞ {am}, which, with their variants ƈɑƞ {8an}, ɑƞ {an} and others, look as if they could map respectively to two of the most common words in *Cantigas d'Amigo*: "meu" ("my") and "eu" ("I"). If that concordance held, we would have a crib for three of the most common glyphs:

- An initial ƈ {8} would represent an initial "m"
- An internal or initial ɑ {a} would represent an internal or initial "e"
- A final ⁀ƞ {m} or ƞ {n} (and these glyphs are nearly always final) would represent a final "u"

Table 12.4. Voynich manuscript: some common "words", and possible counterparts in *Cantigas d'Amigo*

Voynich manuscript					*Cantigas d'Amigo*			
"word"	v101	v104	v104 rank	v104 count	possible counterpart	rank	count	English equivalent
ƈɑ⁀ƞ	8am	8aîΩ	1	851	meu	7	1,058	my
ɑ⁀ƞ	am	aîΩ	3	739	eu	4	1,390	I
oɀ	oe	oe	2	760	se?	6	1,102	himself, herself
					de?	9	991	of, from
ɑʏ	ay	aΨ	5	514	ei	14	775	I have
ɑɀ	ae	ae	12	366				
oʏ	oy	oΨ	6	501	o	8	1,050	the (m.)
o	o	o	25	211				
ƨ	s	s	8	447	e	2	2,432	and
ʏ	y	ψ	16	324				
ɀ	e	e	17	270				

⟃o9	1oe	1oe	9	405	que	1	2,740	that, what
⟃́o9	2oe	1'oe	31	184				
⟃c9	1c9	1cΦ	10	384				
⟃́c9	2c9	1'cΦ	15	298				
⟃c89	1c89	1c8Φ	4	558	quer-	-	84	(I, he, she) want, desire
⟃́c89	2c89	1'c8Φ	7	463				

Source: author's analysis based on the Voynich manuscript, v104 transliteration, and Cohen, Rip, *Cantigas d'Amigo - A Critical Edition* (2010). Voynich "word" counts are based on author's v104 transliteration to Microsoft Word.

If we are feeling lucky, we might even see visual resemblances in this concordance: perhaps the 8 {8} is an "m" rotated counterclockwise, the a {a} can pass for a Latin "e," and the ⌇⌇ {m} and ⌇ {n} are ornate versions of the familiar "u."

The rest of the table gives us some possibilities of extending our crib, as follows:

- A final 9 {e} has a good chance of being a final "e."
- A final y {y} has a good chance of being a final "i."
- An o {o}, possibly in any position, could represent the Galician "o," and in the ubiquitous initial position, could be the Galician masculine definite article "o" ("the"), as in "o fillo" ("the son").
- A ⟃ {1} or ⟃́ {2}, which we suspect are interchangeable, is a good candidate to be "qu" in the ubiquitous Galician word "que."
- An internal 8 {8} could possibly represent the "r" in Galician words of love, such as "quer" ("he/she wants" or "he/she loves"), "quer'eu," "quero" (variations of "I want" or "I love"), "queredes" ("you want" or "you love"), and "queria" ("I/he/she wanted" or "I/he/she loved").
- The ubiquitous final 9 {9} can possibly represent any final vowel; a speaker of any Romance language would know from

the context what the vowel was supposed to be (as in Italian, where "-a," "-e," "-o," and "-I" are endings representing gender and number).

- The glyphs ƨ {s}, ƴ {y} and ɋ {e}, occurring as single-glyph "words," could be variant representations of the Galician word "e" ("and").

We have to admit that some common words from the *Cantigas* are missing from the Voynich top twenty: notably, "non," "por," "ben," "vos," and "ca," as well as "amigo." We could hazard a guess that, if some variant of *Cantigas* is the substrate of the Voynich manuscript, words such as these are abbreviated. Another possibility is that the Voynich manuscript attaches short words to other words. The manuscript has pervasive signs of concatenations of words, and we could well imagine that a phrase such as "meu amigo" ("my friend"), which occurs 556 times in *Cantigas d'Amigo*, becomes a single word "meuamigo," in which case we might be searching for an eight-glyph string beginning with ȣɑ ıı ƞ {8am}. Yet another possibility is that the Voynich manuscript is derived from Galician-Portuguese but from another corpus, with a different vocabulary to that of the *Cantigas d'Amigo*.

The following is our first proposal for what the Voynich scribes might have done with the rest of the common letters in the *Cantigas*, depending on their positions:

- The most common letter in the *Cantigas*, "e," is represented by the Voynich glyph ɑ {a} in the internal and initial positions.
- The letter "a" is possibly represented by some permutation of the glyphs o {o} and ୨ {9}.
- The letter "r" looks like it can be represented sometimes by ȣ {8}, but ᚻ {h} is an alternative which fits well with its frequencies.
- The letter "s" has several possible candidates, depending on the position, including the pervasive o {o} as well as ȣ {8} and ᚻ {h}.

- The letter “n,” in the internal and initial positions, is well served in terms of frequencies by {k}.

And so on down the frequency table.

Table 12.5. The most common letters in *Cantigas d’Amigo* in order of frequency, and possible counterparts among the Voynich glyphs

Galician letter	frequency in Cantigas d’Amigo		examples of lower-case handwriting	expected count in v104 transliteration	visually similar glyphs			
total	position	count		168,262	glyph	v104	position	count
E	internal	15,805		*13,370*		c	internal	14,675
	initial	3,880		*3,282*		c ĉ e	initial	1,690
	final	7,246		*6,130*		c e	final	6,555
A	internal	10,845		*9,174*		a	internal	11,727
	initial	3,713		*3,141*		a	initial	2,724
	final	5,289		*4,474*		a	final	65
O	internal	11,312		*9,569*		o	internal	10,248
	initial	1,144		*968*		o	initial	9,075
	final	4,990		*4,221*		o	final	1,643
R	internal	7572		*6,405*		h	internal	9,233
	initial	593		*502*		h	initial	1,588
	final	5,870		*4,966*		h Ψ	final	6,323
S	internal	4,405		*3,726*		8	internal	8,551
	initial	3,632		*3,072*		8	initial	3,872
	final	5,407		*4,574*		8	final	785

N	internal	3,858		*3,264*		k	internal	5,439
	initial	2,911		*2,462*		k	initial	1,535
	final	6,342		*5,365*		k	final	96
I	internal	7,580		*6,412*		i î	internal	7,192
	initial	375		*317*		i î	initial	9
	final	4,265		*3,608*		i î	final	7
U	internal	8,098		*6,850*		ĉ	internal	2,877
	initial	229		*194*		ĉ	initial	41
	final	3,254		*2,753*		ĉ îΩ	final	4,271
M	internal	2,998		*2,536*		îΩ	internal	40
	initial	7,080		*5,989*		îΩ	initial	2
Q	internal	400		*338*		1	internal	7,999
	initial	4,432		*3,749*		1	initial	9,743

Source: author's analysis. Handwriting examples from Lopes, Graça Videira; Ferreira, Manuel Pedro et al. (2011-) and Mariño Paz, Prof. Ramón: *Edición Cantiga Ajuda, Ejercicios de Lengua y Literatura Gallega*.

actual v104 count substantially above expected count

actual v104 count substantially below expected count

In Table 12.5, we summarized the most plausible mappings that were consistent with the cribs mentioned above and with the frequencies of the respective Galician letters and Voynich glyphs.

A special mention is required for the letter "q," which has a frequency of 2.4 percent in the *Cantigas* and almost always is located at the beginning of a word. It makes appearances in common words such as "quero," "quere," and "queria." We have seen that the glyphs {1} and {2} are plausible candidates for an initial "q." We wonder whether the superscript, which is

always above the pedestal but can be centered, left or right of it, has a function. We might advance the hypothesis that:

- If the superscript resembles ꝛ {y}, as in {2}, then the superscript is an “r” and a bridging vowel “e” is implied, giving us the verb root “quer-“ (“to love, desire”).
- The variant {3}, with a superscript which resembles a final ꝯ {9}, may represent “qua-” as in Galician words such as “quando” (“when”) with 166 occurrences in *Cantigas d’Amigo*, and “quanto” (“how much”) with 131 occurrences.
- The pedestal for a “gallows glyph” represents the letters “qu,” followed by a vowel to bridge to the vertical part of the glyph: for example if, as seems possible, {h} can represent “s,” then {H} can represent “quis-“ as in the Galician words “quis” (I/he/she loved”) or “quiser” (“that I/he/she will love”).

Finally, we shall explore the possibility that {k} can sometimes (maybe not always) represent the digram “fr,” which can be written in Galician as follows: .

With that, let us put our mapping to the test.

VOYNICH TO GALICIAN-PORTUGUESE

We return to page f001r, the opening page of the Voynich manuscript in its current binding. This was the first page when Wilfrid Voynich purchased the manuscript; we know that at that time, some pages were missing and that the manuscript had been rebound at least once. We do not know whether f001r was the first page when the manuscript was first written, but we may imagine that it was. If so, surely we could expect it give some sign of what the book is all about. If the manuscript has any connection with

medieval Galician-Portuguese or with *Cantigas*, we suppose that this page is as good a place to start as any.

Specifically we propose to examine the second paragraph of this page. This paragraph contains four lines (numbered 07 to 10 in the v101 transliteration). Line 07 begins with the "rare glyph" that we mentioned in Chapter 4. This glyph resembles a bird on the wing, or the antlers of a moose or deer; but we might conjecture that it is an initial letter, embellished in this instance to the point of incomprehensibility. Line 10 is aligned vertically with Line 09, but is approximately right-justified. Lines 9 and 10 end with the same three glyphs; we might say that they rhyme. Figure 12.4 reproduces the paragraph in question.

For sure, most of it does not look much like Galician-Portuguese. But on the last line, do we see the phrase "senhor fremosa"?

For if we do, we have come across one of the most-common phrases in *Cantigas*: one which occurs, with variations, 132 times.[9] Here are a few examples, all from the *Cantigas de Amor*:

> senhor fremosa, a mim a guisou (by Pero de Armea)
> senhor fremosa, ca vós me matades (by Pero da Ponte)
> senhor fremosa, ca per vós perdi (by Pero Garcia Burgalês)
> senhor fremosa, des que vos amei (by Juião Bolseiro).

The feminine adjective "fremosa" means "beautiful." The songs quoted above are clearly addressed to a woman, and the salutation "senhor" implies "senhora" ("Lady").

Perhaps we have discovered a line in the Voynich manuscript which reads "beautiful Lady."

Let us then return to the glyphs in their mute elegance. Let us consider that some of them may be abbreviations; that some of them may have more than one interpretation; and that some of them may be plain errors of handwriting, that the scribes were never asked to correct (perhaps because the authors or producers never intended the work to be understood).

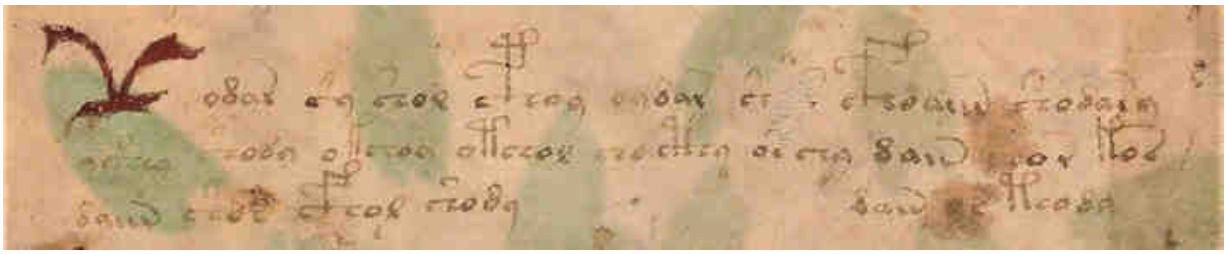


Figure 12.4. Voynich manuscript, extract from page f001r, lines 07–10. *Beinecke Rare Book and Manuscript Library*

On that basis, here is this author's free transliteration of the last two lines into medieval Galician-Portuguese:

Line 9	sempre Groriosa
Line 10	Senhor fremosa

We may offer the perception that, if there is any narrative meaning whatsoever in this transliteration, it represents a poem or a song about love and it is addressed to a beautiful lady.

What is needed is a task of trial and error, involving alternative interpretations of the glyphs; a deeper search for abbreviations, misspellings, and omissions; and a sense of the literary context. Here, we can only solicit the knowledge and experience of readers who have a deep and intuitive knowledge of Galician-Portuguese.

This author reached out to professors of Galician-Portuguese in the great universities of Galicia, including those of Vigo, Coruña, and Santiago de Compostela, as well as to scholars elsewhere in Spain and in Canada and Portugal. None had the time or the inclination to participate in this project. This chapter, therefore, simply represents an example of this author's best efforts to extract a Galician-Portuguese meaning from the Voynich manuscript and stands also as an invitation to readers of this book, to continue the task.

CHAPTER 13

ILLYRIA

From our prologue we recall that around the fall of 1637, Athanasius Kircher, the celebrated professor at the Collegium Romanum, received from Theodorus Moretus a hand-copied sample from an old book. It has been claimed that this book was what we now call the Voynich manuscript (although for all that we know, it could have been another book altogether). Kircher replied to Moretus on March 12, 1639, to the effect that he saw no intrinsic difficulty in deciphering it (if that is the right word), and implied that he would get round to it when the mood took him.

Kircher's letter is preserved to this day at the Czech National Library, and the critical extract is reproduced in Figure 13.1:

The Latin appears to read as follows:

> Alterum denique folium quod ipsi ignoto characteris genere scriptum uidebatur illyrico idiomate, charactere quem D. Hieronymi uulgo uocant, impressum sciat; utunturque eodem

charactere hic Romae in missalibus alijsque sacris libris illyrico sermone imprimendis.[1]

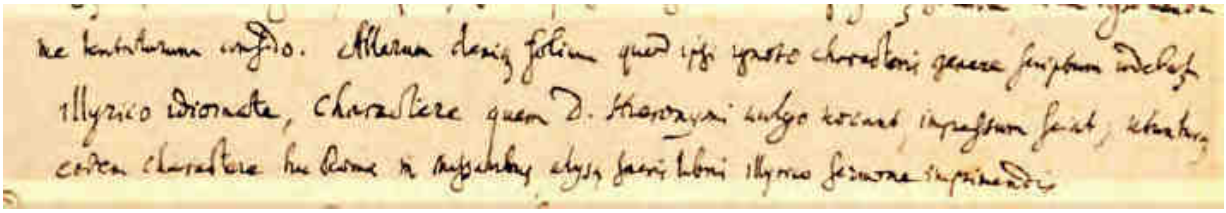


Figure 13.1. Extract from the letter of Athanasius Kircher to Theodorus Moretus, dated March 12, 1639. *Czech National Library*

This author's translation to English is, approximately:

> Finally, the other sheet that seemed to him written in a kind of unknown characters in the Illyrian language, let him know that it was printed in the characters that people call those of D. Hieronymus; and they use the same characters here in Rome to print missals and other sacred books in Illyrian.[2]

The word "ipsi," meaning "to him," refers to a third person previously mentioned. There is no other person mentioned in Kircher's letter. We could suppose that Moretus, in his original letter to which the writing sample had been attached, had mentioned Barschius as the contributor of the sample. If so, perhaps Barschius had raised the idea that the language was Illyrian, and Kircher was asking Moretus to pass on a message to Barschius.

To this author, it is clear that Kircher identified the script of the second sheet, but it is not clear that he identified the language. If we in modern times were to see a text in Cyrillic script, we would not necessarily know whether the language was Russian, Ukrainian, Bulgarian, or some other Slavic language. Kircher implied simply that the script was known and that it was used in documents in the Illyrian language. That would not exclude the use of the same script in other languages. Finally, it is unclear whether either Moretus or Kircher thought that the first and second sheets were in the same script.

Kircher was a professor of mathematics, physics, and Oriental languages. He has been described as "the last man who knew everything."[3]

We can surely believe that whatever his observations meant, he did not make them lightly.

Has the language of the Voynich manuscript been hiding in plain sight for nearly five hundred years?

WHERE WAS ILLYRIA?

But first: what did Kircher mean by Illyria?

In Kircher's time, there was no state, nation or province bearing the name of Illyria. No person on the streets of Rome, or for that matter of any European city, would have identified himself or herself as an Illyrian. Not until October 14, 1809, over 150 years after Kircher's letter, would there be an administrative region bearing the name of Illyria; and this would be when the First French Empire created the Illyrian Provinces from the territories of the defeated Austrian Empire. These provinces incorporated the former Austrian (and some previously Venetian) lands of western Carinthia, Carniola, Dalmatia, Gorizia and Gradisca, Istria, Lienz, Ragusa, and Trieste. Their boundaries, approximating those of modern Croatia, are shown in Figure 13.2.

Figure 13.2. The Illyrian Provinces of the First French Empire, as created on October 14, 1809. *Andrein, Ziegelbrenner*

Kircher could only have been referring to a more ancient Illyria: in fact, to an Illyria, which had probably ceased to exist as a political entity more than 1,200 years before his time.

For an understanding of the genesis of classical Illyria, we are indebted to the authoritative work of John Wilkes, professor of Greek and Roman Archaeology at University College London.[4]

Here, from the pen of the Greek historian Appianus Alexandrinus, or Appian of Alexandria, (ca. 95–ca. 165) is a Greek perception of the boundaries of Illyria:

> The Greeks call those people Illyrian who dwell beyond [from the Greek perspective: north of] Macedonia and Thrace, from Chaonia and Thesprotia to the river Danube. That is the [north–south?] length of the country, while its [east–west?] breadth is from Macedonia and the mountains of Thrace to Pannonia and the Adriatic and the foothills of the Alps. It is five days journey in breadth and thirty across, say the Greeks. The Romans measured the country as above six thousand stades in length (750 Roman miles)[5] and about twelve hundred stades (150 Roman miles) in width.[6]

The Greek description of the Illyrian lands seems to encompass most of modern Albania, Bosnia & Herzegovina, Croatia, Kosovo, Montenegro, North Macedonia, Serbia, and Slovenia.

Appianus was writing about a people who had been known to the Greeks for at least four centuries. An unnamed mariner wrote in the fourth century BC about a circumnavigation that he had made of the Adriatic coast. He had traveled in a clockwise direction: that is, southward along the eastern shore, and previously (or so we assume, not having access to the original document) northward along the western shore, the coast of what is now Italy. His memoir was titled *Periplus*, or *The Coastal Passage*. His is the earliest known reference to the Illyrians.

It is clear that the unknown sailor started his southbound leg around what is now Trieste and passed the peninsula which in our time bears the name of Istria and is mainly part of modern Croatia. This sets an approximate northern boundary to the Illyrian lands.

> After the Veneti the river Ister…. The coastal voyage along the Istrian region lasts a day and a night…. After the Istri is the people of the Liburni….

The sailor took another two days to pass the lands of the Liburni. At this point for the first time, he encountered the Illyrians:

> … After the Liburni there come the lllyrian people. The Illyrii dwell by the sea as far as Chaonia, which lies opposite Corcyra, the island of Alcinous. There is situated the Greek city called Heraclea, with a harbour…. The Bulini are also an Illyrian people…. the trading settlement [on the River Naron] … lies upwards of eighty stades from the sea. The people living there are the Manii, who are by race Illyrians…. Beyond, there is a huge lake, extending from the inland side of the trading settlement as far as the Autariatae, an Illyrian people…. The Enchelei are an Illyrian people, who inhabit the land after Rhizon…. The Taulantii are an Illyrian people, in whose land is the city Epidamnus…. Then from Epidamnus to Apollonia, a Greek city, the journey on foot takes two days…. From Amantia more within the Ionian Gulf is the city Oricus…. These are the Illyrian peoples, extending from the Bulini up to this point.[7]

Corcyra is the modern island of Corfu, and Chaonia, the mainland to the north of Corfu, is now part of southern Albania. This then was the southern boundary of the Illyrian lands.

The sailor recalled a sequence of settlements, which he described as sometimes Greek and sometimes Illyrian, although the inhabitants of each settlement had their own name for who they were. We can surely assume that he based his descriptions on the languages that he heard; therefore, that he understood both Greek and Illyrian. Other than these two races, if we may call them that, he did not mention any other ethnic group in his southbound journey.

In respect of the Illyrians, he seemed to be describing a race of people who, interspersed and coexisting with Greek settlements, inhabited the

eastern coast and hinterlands of the Adriatic Sea, from somewhere south of modern Istria to somewhere north of modern Corfu.

To make an analogy from another time and place: in the ninth century AD, along the east coast of what is now England, a sailor traveling southward would have encountered a succession of settlements whose names are now lost to us, but whom we could characterize today as Northumbrians (whom the sailor would have described as of the English race), Mercians (Danes), East Anglians (Danes), and the Kentish people (English). We do not know how these people called themselves at that time, but our hypothetical sailor would have heard their language and would have remembered them as Danes or else English.

In the map in Figure 13.3, adapted from the work of Professor Wilkes, we illustrate the extent of the Illyrian lands as they were described by Appianus Alexandrinus and implied by the unknown navigator of the Adriatic Sea.

At some point between the fifth and third centuries BC, according to Professor Wilkes, some of the Illyrians metamorphosed from a group of trading settlements into a political power (possibly a federation) with a navy and an army, capable of making war on the Macedonians and Epireans to their south. Professor Wilkes describes the process:

> From the end of the fifth century [BC] we have seen how it was possible for regimes among the southern lllyrians to exploit periods of weakness in Macedonia and Epirus, notably Bardylis of the Dardani and Glaucias of the Taulantii…. There is now [in the third century BC] an impression, although it can be no more than that, of a new and more lasting political order emerging among the lllyrians, not only among the Dardanians. … For around a century the kingdom of the lllyrians was a minor but stable power on the margin of the Hellenistic world.

Figure 13.3. The approximate boundaries of the Illyrian lands between the fourth century BC and the first century AD. Base map from USGS. *Graphics by author, adapted from Wilkes, John (1995)*, The Illyrians, *Map 1*

We might perhaps more accurately speak of the Kingdoms of the Illyrians, for the political order in question consisted demographically of at least fifteen distinct Illyrian peoples, each with their own king, and geographically of a specific group of settlements, dispersed throughout what is today Albania.

The territory of the Illyrian kingdoms is illustrated in Figure 13.4.

In any event, the kingdoms of the Illyrians would come to an end at the hands of the Romans. Between 229 BC and 168 BC, the Roman legions made war three times on Illyrian forces. The stories of the three Illyrian wars have been well told elsewhere, notably by John Wilkes. For our purposes we may record that in 168 BC, the southern Illyrian kingdoms were defeated, their alliances were dissolved, and their lands were transformed into three provinces under Roman control. Over the next century and a half, further Roman campaigns subdued classical Illyria. By 27 BC, the historical lands of Illyria had been absorbed into the Roman provinces of Dalmatia and Pannonia, collectively known as Illyricum.

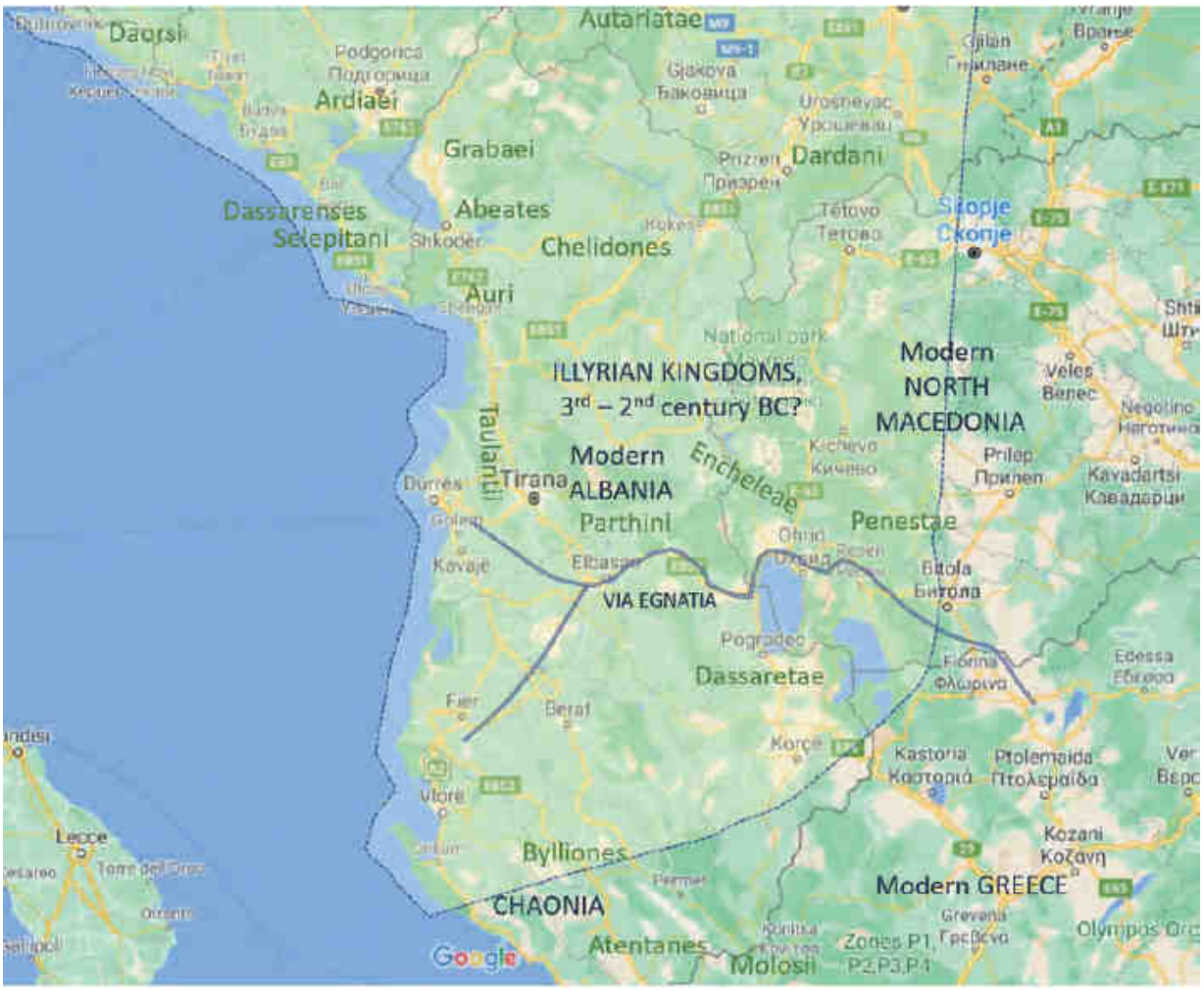


Figure 13.4. The approximate locations of the Illyrian kingdoms between the third and second centuries BC. Base map © 2021 Google Maps. *Graphics by author, adapted from Wilkes, John (1995)*, The Illyrians, *Map 3*

Thus, would the Illyrian lands remain until near the close of the fourth century AD. John Wilkes describes the end of an era:

> Roman Illyricum ended on 9 August AD 378, when the emperor Valens and his armies were wiped out by the Goths in a battle at Adrianople in southeast Thrace…. for Illyricum as a whole the invasions at the end of the fourth century destroyed not only the structure of the Roman province but possibly a good deal of the older native lllyrian cultures which had survived within it.

With that, for all practical purposes, Illyria disappeared from history.

Nevertheless, Athanasius Kircher, writing in the seventeenth century, had a clear idea in his mind of where Illyria was. In *Turris Babel* (*The Tower Of Babel*), published in 1679, he wrote:

> "Illyria [is] today Slavonia, or Dalmatia, in ancient times Thrace, called after Thyras the descendant of Japheth, who was the first possessor of this country …
>
> … although to this day, what was formerly called Thracia, or Zervia, has been divided into innumerable regions … such as the Zervi, the Dalmatians, the Bosni, the Croatae, the Rasci, the Valdavi, the Walachi, the Podoli, the Dacians, and the colonies of the Boemi, the Moravians, the Poles, and the Muscovites …[8]

In other words, Kircher's geographical expression of Illyria seemed to refer to most of the Balkan peninsula north of Greece, apparently including present-day Bosnia & Herzegovina, Croatia, and Serbia. His references to the Dacians, Podoli, Valdavi (Moldavians?), and Walachi seemed to cast his Illyrian net as far east as present-day Romania, Moldova, and Ukraine. However, he did not mention Rumelia (which at his time of his witing encompassed most of what is now Albania, Bulgaria, northern Greece, Kosovo, and North Macedonia).

With the exception of several coastal strips of Dalmatia which were under the Venetian Empire, and the Habsburg domains in parts of present-day Croatia and Slovenia, most of the lands that Kircher classified as Illyrian were under the direct or indirect control of the Ottomans. The Muslim rulers and administrators of the Ottoman Empire granted freedom of religion to their non-Muslim subjects. Kircher would have known that the majority of the residents of the Illyrian lands were Christian, predominantly adherents of the Eastern Orthodox churches. Still, there were sufficient numbers of Catholics in Illyria for Rome to take an interest in them.

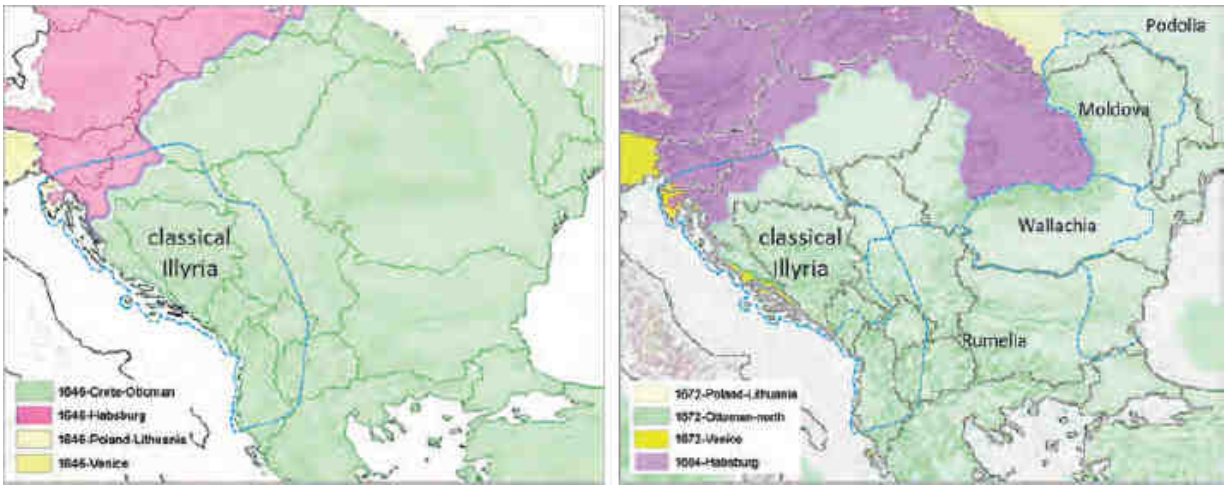


Figure 13.5. The boundaries of the Hapsburg, Ottoman, and Venetian Empires on the Balkan Peninsula in 1646 and 1672, and of classical Illyria in the fourth to first centuries BC. *Base maps by courtesy of Dr Bogdan Gabriel Popescu and Professor John Wilkes; additional graphics by author*

THE ILLYRIAN LANGUAGE

John Wilkes has this to say about the Illyrian language:

> [The Illyrians] spoke a language of which almost no trace has survived. That it belonged to the 'family' of Indo-European languages has been deduced from the many names of Illyrian peoples and places preserved in Greek and Latin records, both literary and epigraphic. We cannot be sure that any of them actually called themselves Illyrians: in the case of most of them it is near certain that they did not.

What then are we to make of the statement of Athanasius Kircher in 1639, over 1,200 years after Illyria had ceased to exist, that the Illyrian language was still in use as a vehicle for sacred writings?

Here we need to introduce Eusebius Sophronius Hieronymus the Illyrian, better known as St. Jerome, whom Kircher mentioned in his letter to Moretus.

Hieronymus was born around 345 AD in a settlement called Stridon, which no longer exists. The most specific identification of this place is that of William R. Shepherd, whose *Historical Atlas*, dated 1911,[9] recorded Stridon as the seat of a Catholic bishopric in the Roman province of Dalmatia. That location is within modern Bosnia & Herzegovina. Hieronymus was a prolific writer on theology and created an authoritative Latin translation of the Bible, now known as the Vulgate. He died around 420 AD and was canonized as Saint Jerome in 1767.

Kircher's reference to Hieronymus (who was not yet known as Jerome) must surely have arisen from a tradition that Hieronymus was the inventor of a specific script for writing the Illyrian language.

Among the documents that supported this tradition, Kircher must have been aware of the following:

- Guillaume Postel (1538): Alphabetum Hieronymianum seu Dalmaticum, aut Illiricum [The Alphabet by Hieronymus of Dalmatia or Illyricum) in *Linguarum duodecim characteribus differentium alphabetum* [*The twelve characters of the alphabet in different languages*]
- Giovanni Batista Palatino (1545): Alph. Illiricum divi Hieronymi [The Illyrian Alphabet of St. Hieronymus] in *Libro Nuovo* (*Libro nel qual s'insegna a scrivere ogni sorte lettera, antica et moderna*) [*New Book (Book in which it is shown how to write all kinds of letters, ancient and modern)*]
- Angelo Rocha (1591); Alphabetum Illiricum [the Illyrian Alphabet], (which he attributed to "holy Hieronymus the Illyrian") in *Biblioteca Apostolica Vaticana a Sixto V Pont. Max*. [*Vatican Apostolic Library of Sixto V the Supreme Pontiff*]
- Claude Duret (1619): Alphab. Illyricum divi Hieronymi [Illyrian alphabet of St. Hieronymus], in *Thresor de l'histoire des langues de cest Univers* [*Treasury of the history of languages of this universe*].

This tradition would continue for at least a century after Kircher's time, as evidenced by the inclusion of the Illyrian alphabet, still attributed to

Hieronymus, in the following documents:

- Denis Diderot and Jean le Rond d'Alembert, eds. (1751): Alphabets Orientaux Modernes: Illirien ou Hieronimite [Modern Eastern Alphabets: Illyrian or Hieronymite], in *Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers* [*Encyclopedia, or a Systematic Dictionary of the Sciences, Arts, and Crafts*]
- Pierre Simon Fournier (Paris, 1766): Alphabets: [number] 68 Illyrien de S. Jerôme, in *Manuel typographique, utile aux gens de lettres, & à ceux qui exercent les differentes parties de l'Art de l'Imprimerie* [*Typographic manual, useful for writers and for those who exercise the different parts of the art of printing*].

We have summarized the Illyrian alphabet, as presented in these six documents, in Figure 13.6.

From this figure, we can make the following observations:

- There is practically universal agreement between the six documents, spanning over two centuries.
- The shapes, names and pronunciations of the Illyrian letters did not change significantly during this period.
- Only in Diderot and D'Alembert's alphabet was there a distinction between uppercase and lowercase letters, and practically always, the lowercase was just a miniature version of the uppercase.

We may also observe that Hieronymus' script (if indeed it is his) is not an efficient mechanism for writing a language. Most of the characters need a multitude of pen or quill strokes. Even the vowels— ꟸ[a], ɜ [e], ɤ [i], ꟹ [o], ꟻ[u] —which in most phonetic languages are the most frequent letters, do not have the simplest of shapes and look as if they would be laborious to commit to paper. The letters do not lend themselves to cursive writing. Only Diderot and D'Alembert attempted to write the names of the letters in the

script itself (for example they wrote “glagole” as ⰳⰾⰰⰳⰻⰻⰵ; but they made no attempt to join up the letters.

The Hieronymus alphabet lives on. Better known now as the Glagolitic alphabet, it occupies a block in the modern Unicode character set. The Unicode symbols for the Hieronymus letters from our six sources (plus a few others not listed in those documents) are included in Figure 13.6 alongside the historical representations.

We must however return to Kircher’s writings for an idea of what he meant by the Illyrian language (as distinct from the Hieronymus script, which is applicable to any phonetic language). We recall that Kircher was a professor of Oriental languages at the Collegium Romanum. In *Turris Babel*, Kircher wrote as follows:

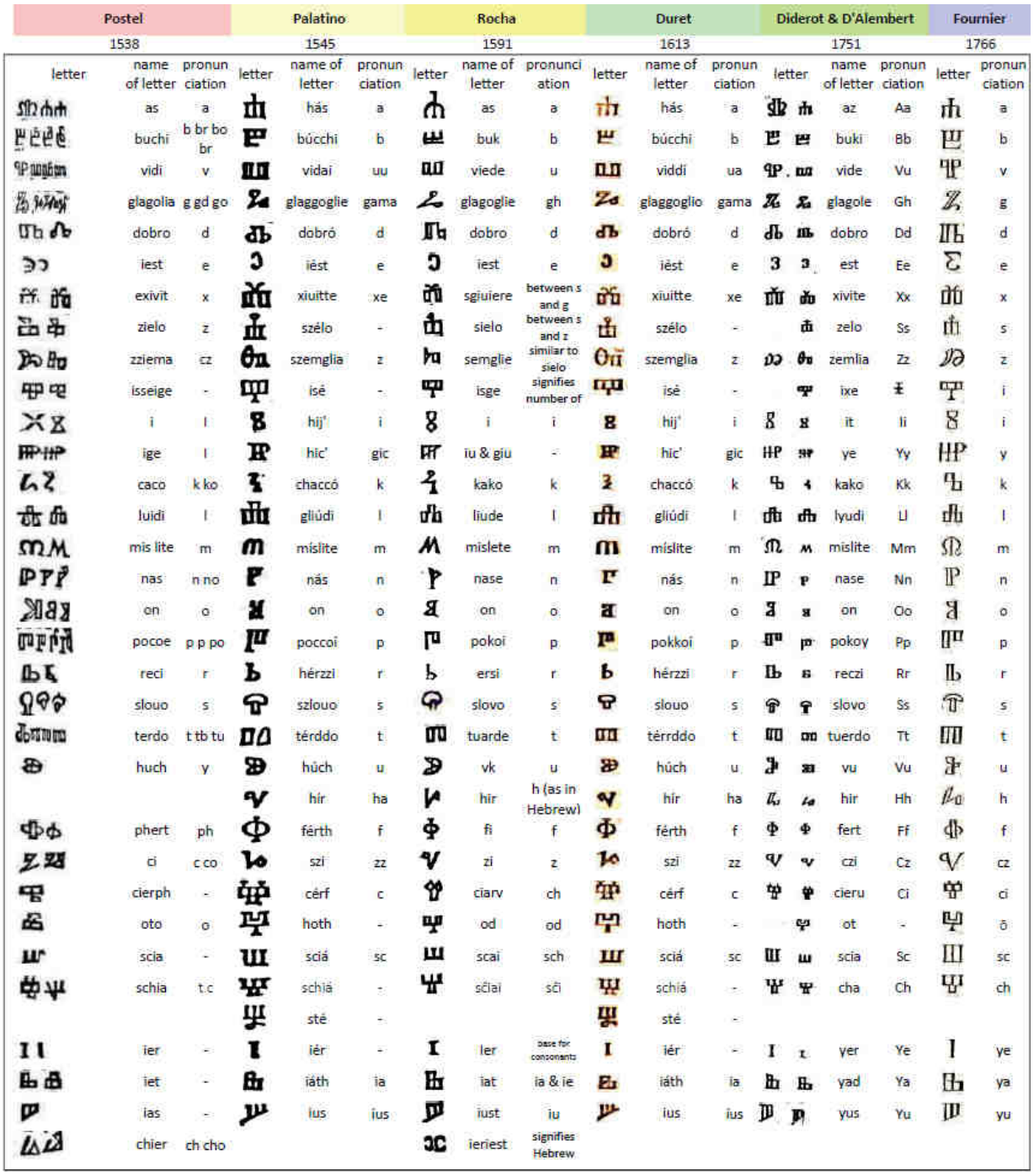


Figure 13.6. The Illyrian alphabet, attributed to Eusebius Sophronius Hieronymus, as presented in selected documents from the sixteenth, seventeenth, and eighteenth centuries, and the modern Unicode symbols for the Glagolitic alphabet. *Sources: Postel (1538), Palatino (1545), Rocha (1591), Duret (1613), Diderot & D'Alembert (1751), Fournies (1766)*.

There remains Illyrian, the most ancient language, which has transmitted its branches into Poland, Lithuania, Moscow,

> Tartaria, and even to the farthest parts of the East; it was translated into innumerable languages; from this, the Tunic, Bulgarian, Bosnian, Tartar filled up a very great portion of Tartaria. But they are so corrupted that they retain scarcely any trace of the primitive….[10]
>
> … It has, however, several dialects peculiar to different nations; … all the countries under Thracia or Illyria use this one language … it is clear that the Illyrian language, or Slavonic, Bohemian, Polish, Lithuanian, Muscovite, and what is mainly in use in the Court of Constantinople, is one and the same language, only different, because of the condition of the different regions, and the corrupted dialects, corruption being common to all languages over time.[11]

In effect, Kircher was saying that Illyrian was the ancestor of all the Slavic languages. In this respect, his view was diametrically opposed to that of most modern scholars, whose consensus is that the Slavs arrived in the Balkan peninsula in the seventh century, displacing the indigenous people and cultures. An analogy, we suppose, would be the Anglo-Saxon displacement of the indigenous Welsh, Scots, and Cornish, and of their languages, in what is now Great Britain. By implication, the modern view is that Illyrian has no affinity with the Slavic languages.

Kircher went on to make another extraordinary statement:

> The Illyrians have a twofold alphabet of their own letters, which they use in writing books.[12]

Forty years earlier, in 1639, he had written to Moretus to say that missals (and he could only have meant Roman Catholic missals) were written in Rome in the Illyrian language with the Hieronymus script. In 1679 he wrote that the Illyrians themselves wrote books in their own alphabet; could he have meant anything other than the Hieronymus script? Yet, modern scholars would have us believe that, of all the missals written in Rome for the Illyrians, and of all the books written by the Illyrians themselves (surely in both cases, over several centuries), nothing remains.

Finally, we have to consider whether what Kircher saw in the second sheet of Moretus's letter was indeed the Hieronymus script but applied to what we would now call a Slavic language. From his later writings, we know that Kircher ultimately came to think of all the Balkan languages as one and the same, but all with Illyrian as their primitive ancestor. In 1639, when he wrote to Moretus, he was just thirty-six years old; we do not know whether at that time he thought of Illyrian as a distinct language.

It is clear that the Illyrian language and the Hieronymus alphabet were well known to Athanasius Kircher. From Kircher's letter of 1639 to Moretus, we can reasonably infer that, at least in Kircher's time, there were sufficient numbers of Roman Catholic believers within the Illyrian lands, to make it worthwhile to translate and print holy books in their language.

This being so, we can attempt some rough and ready demographics.

DEMOGRAPHICS OF THE ILLYRIANS

J. C. Russell estimated that in 350 AD, shortly before the Roman province of Illyricum came to an end, the population of the European part of the Roman Empire was about 18.3 million.[13] The area of the Roman Empire, at its peak under the emperor Trajan (who ruled from 98 to 117 AD), was about 5 million square kilometers.[14] We may estimate the European part of the Empire at about 3 million square kilometers. The average population density in Roman Europe was therefore about six persons per square kilometer.

The Illyrian lands, at their greatest extent, had occupied what most of what are now the southern Slav states (the former components of Yugoslavia) plus most of what is now Albania. The area of these modern states is 285,000 square kilometers; we may adopt a figure of about 250,000 square kilometers as an estimate of the area of the Illyrian lands.

Here we have to take into consideration that the Illyrian lands were largely mountainous and could not have supported a population density comparable with that of Roman Europe as a whole. Let us make the heroic assumption that the Illyrian lands had a population density of three persons per square kilometer (half the European average). If so, we can estimate

that shortly before their province fell to the Goths, the Illyrian lands had a population of about 750,000. Some part of this population consisted of native speakers of Greek, Macedonian, Latin, or other languages of the Roman Empire. Our figure places an approximate upper limit on the number of native Illyrian speakers as of the year 350 AD.

We fast-forward to 1639, when Kircher wrote his letter to Moretus. Jan de Vries has estimated that in 1650 the population of Europe, excluding European Russia and Turkey, was about seventy-five million,[15] in other words about four times that of Roman Europe at the time of the fall of Illyricum. The Illyrian language had survived, as Kircher assures us. Then we could reasonably suppose that the Illyrian speakers, even without a state or province of their own, had kept pace in demographic terms with the population of Europe.

If all these assumptions bear scrutiny, we could imagine that in Kircher's time there were up to three million Illyrian speakers who were prospective readers of holy scripture. We can readily understand Kircher's observation that "they use the same script here in Rome to print missals and other holy books in the Illyrian language."[16]

At some point, Hieronymus's script would acquire another name: the Glagolitic script (after the fourth letter in the script, "glagolia"), and would be used to write sacred and profane texts in other languages, notably the southern Slavic languages. But of the texts written in this script in the Illyrian language, of which Kircher wrote so authoritatively, we are told that no trace remains.

ILLYRIAN AND VOYNICH

We may now turn to the issue of whether there is any relationship between Hieronymus' characters and the Voynich glyphs, as Kircher's letter permits us to conjecture.

We can think of at least two ways to tackle this:

- to return to the "Gold Bug" approach, in which we compare the frequencies of the Illyrian letters with the frequencies of the

Voynich glyphs and see whether we can find some match or correlation, at least for the most common letters or glyphs

- to examine the Illyrian letters and the Voynich glyphs visually for any signs of resemblance

Since resemblance may be in the eye of the beholder, let us start with the "Gold Bug" approach.

THE "GOLD BUG" AGAIN

Here we run into the first roadblock. As we noted earlier: there are no surviving texts in the Illyrian language. For sure, there are many surviving texts in the Hieronymus script, but they are not in the Illyrian language. It is somewhat as if all Latin texts had disappeared, and all that remained were modern European texts written in their respective versions of the Latin script. In that case we would know nothing of letter frequencies in Latin. Likewise, at this point we have no way to calculate letter frequencies in Illyrian.

We can maybe find a way around the roadblock. Let us suppose that the Illyrian language has survived in some form to the present day, and is no longer known as Illyrian, but that its traces or echoes remain in some European language. We might further suppose that the Illyrian speakers did not migrate in their entirety to other parts of Europe, and that some of the descendants of the Illyrians remain on the ancient Illyrian lands. In that case we might search for a descendant of the Illyrian language in the southern Slav states (the former Yugoslavia) and in Albania.

Authors in various Slav states and in Albania, we suppose out of national pride, have claimed that their respective peoples are the descendants of the Illyrians. We have no way to judge these claims. However, the modern scholarly consensus is that the Slav peoples first arrived in the Illyrian lands in the middle of the sixth century AD. So, it seems reasonable to suppose that the Slav languages are not descendants of the Illyrian language. In that case, we could more productively search for echoes of Illyrian among the non-Slavic languages of the former Illyrian lands: of which the principal

one today is Albanian. The modern Albanian language is spoken by about six million people in the Balkans, principally in Albania and Kosovo.[17] Today's Albanian speakers are a people comparable in number with the number of Illyrian speakers that we have imagined in antiquity.

Let us make another heroic assumption: that what Athanasius Kircher saw, attached to the letter of Moretus, was a document in Old Albanian.

The oldest extant document in Old Albanian is *Meshari* (in Latin, *Missale*), written by Gjon Buzuku in 1555. It is a translation of part of the Catholic liturgy into the Gheg dialect of Albanian. By courtesy of the TITUS Project at the University of Frankfurt, we have used the full text of *Meshari* as our corpus for the study of Old Albanian. This text contains, by this author's count, 75,699 words comprising 326,356 characters excluding spaces. The average length of a word in *Meshari* is 4.31 letters (which is not much different from the average length of "words" in the Voynich manuscript).

Apart from punctuation marks, Buzuku used a total of forty-three characters: twenty-four unaccented letters from the Latin alphabet (all except "k" and "w"), fourteen accented or modified Latin letters (but not the "ë," which is common in modern Albanian), and five letters that could be interpreted as Greek or Cyrillic. By comparison the Hieronymus alphabet, in the compilations that we have identified, has thirty-two or thirty-three characters.

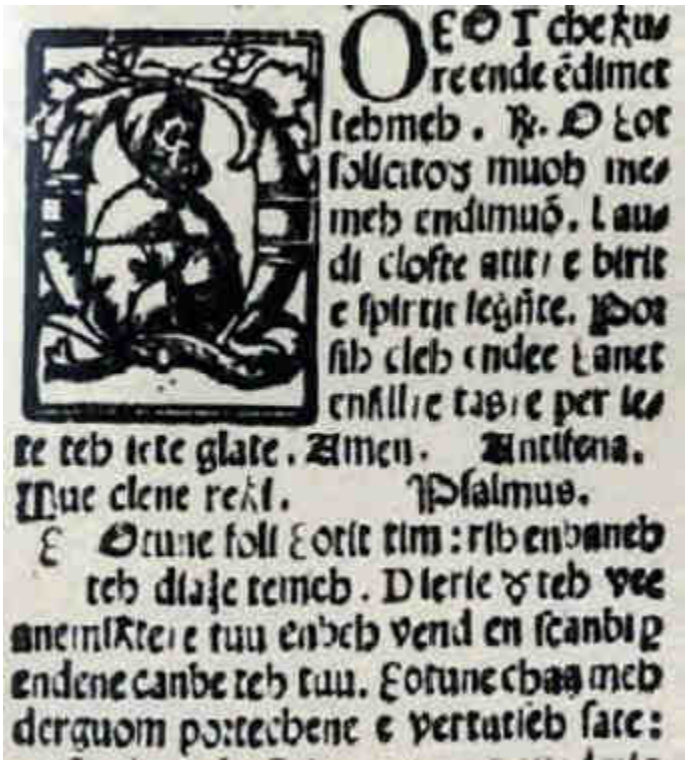


Figure 13.7. An extract from *Meshari* by Gjon Buzuku (1555). *shqipopedia.org*

We could propose that by the year 1555, the ancient Illyrian language had metamorphosed into a dialect of Albanian. We know, thanks to Athanasius Kircher, that the Hieronymus alphabet was used as late as 1639 for writing Catholic missals in what Kircher called "the Illyrian language." We can even suppose that Buzuku, as the author of the first missals in Albanian, knew the Hieronymus script. So we can readily imagine the scribes of the Voynich manuscript working on the transcription from source documents in Old Albanian and in the Hieronymus script, perhaps even Catholic missals with their poetic and repetitive structure.

The most-common letters in *Meshari* are as follows:

> E (21.4%) I (9.2%) T (9.0%) H (8.2%) N (6.9%) A (6.3%).

We may here observe that the ubiquity of the letter "E" in Buzuku's writing evokes the extraordinary frequency of the o [o] in the Voynich manuscript.

If Buzuku or any other medieval Albanian scribe had written in the Hieronymus script, we might have expected his most-common letters to be as follows:

𐓎 = E 𐓘 = I 𐓣𐓣 = T 𐓭𐓱 = H 𐓤 = N” 𐓱𐓱 = A

We may now make comparisons with the most-common glyphs in our v104 transliterations of the Voynich languages; and in order to respect the position-dependency of the Voynich glyphs, we have to make separate comparisons for the initial, internal, final and isolated positions of the letters in Old Albanian words and the glyphs in Voynich “words.” We attempted these comparisons for Language B only, this being the predominant Language of the Voynich manuscript. A summary of the results is presented in Table 13.1. Some observations:

- The frequencies of the Voynich glyphs match quite well with the frequencies of the Old Albanian letters (and by implication, those of our hypothetical Illyrian texts in the Hieronymus script). This is true for the initial, internal, and final positions and for isolated letters (one-letter words).
- The Voynich glyphs do not greatly resemble the Hieronymus letters to which we are proposing that they correspond. Therefore, if the Voynich scribes had a reference chart that mapped Hieronymus characters to glyphs, there was no great effort to make the glyphs similar to the source characters.
- The mapping between letters and glyphs changes depending on the position of the letter. For example, the ubiquitous Old Albanian letter E, if an initial letter of a word, could map to the glyph ꞇo [Δ]; if internal to a word, to the glyph ᴄ [c]; if the final letter of a word, to the glyph ϙ [Φ]; and if a one-letter word, to the glyphs ϙ [e} or ϙ [Φ] interchangeably.

Table 13.1. Voynich manuscript, Language B: possible position-dependent mappings between glyphs, Old Albanian letters and Hieronymus alphabet

	Voynich Language B			Old Albanian		equivalent Hieronymus letter		
glyph	v101	v104	count	letter	count			
isolated letters (words of one letter)				RSQ	83.7%			
ꬹ ୭	e, 9	e, Φ	452	E	5,629	est	Ⱅ	ⱅ
ʔ	y	Ψ	195	Ɣ	1,000	glagole?	Ⰳ	ⰳ
ƨ	s	s	174	Ꝑ	610	?		
ο	o	o	124	O	297	on	Ⰻ	ⰻ
†ο	4o	Δ	120	Ẽ	111	?		
ȣ	8	8	57	A	30	az	Ⰰ	ⰰ
initial letters (words of at least two letters)				RSQ	90.7%			
ο	o	o	5,212	T	8,743	tuerdo	Ⰵ	ⰵ
†ο	4o	Δ	3,960	E	7,128	est	Ⱅ	Ⱅ
ꝛ	1	1	3,619	P	4,835	pokoy	Ⰿ	ⰿ
ɑ	a	a	2,136	M	4,817	mislite	Ⰽ	ⰽ
ꝛ̊	2	2	2,110	ſ	4,233	zelo?	Ⰷ	ⰷ
ȣ	8	8	1,734	C	4,169	cieru	Ⰸ	ⰸ
internal letters (words of at least two letters)				RSQ	92.4%			
ꞇ	c	c	9,735	E	40,714	est	Ⱅ	ⱅ
ɑ	a	a	7,506	I	21,439	it	Ⰺ	ⰺ
ȣ	8	8	6,693	N	17,341	nase	Ⰼ	ⰼ
ᚻ	h	h	5,830	T	15,359	tuerdo	Ⰵ	ⰵ
ο	o	o	3,519	A	13,514	az	Ⰰ	ⰰ
ᚠ	k	k	2,932	U	12,361	vu	Ⱆ	ⰵ
final letters (words of at least two letters)				RSQ	88.8%			
୭	9	Φ	10,418	H	17,356	hir	Ⰴ	ⰴ
ᐁ	N	Ω	3,999	E	12,915	est	Ⱅ	ⱅ
ꬹ	e	e	3,492	I	4,780	it	Ⰺ	ⰺ
ʔ	y	Ψ	3,409	T	3,852	tuerdo	Ⰵ	ⰵ
ο	o	o	630	R	2,223	reczi	Ⰱ	ⰱ
ȣ	*	Π	618	A	2,082	az	Ⰰ	ⰰ

Source: author's analysis. Letter frequencies in Old Albanian based on Buzuku, Gjon: *Meshari* (1555). Note: The table shows only the top six glyphs and letters in each position. Correlation coefficients are based on the 26 glyphs in the v104 alphabet and on the top 26 letters in Old Albanian, in each position.

Table 13.2. Voynich manuscript, Language B, v104 transliteration: most common "words" of two, three, four and five glyphs, and transliterations to Old Albanian based on position-dependent glyph and letter frequencies

Voynich		Old Albanian	occurrences in *Meshari*		Voynich		Old Albanian	occurrences in *Meshari*	
glyphs	v104		as strings	as words	glyphs	v104		as strings	as words
two-glyph "words"					three-glyph "words"				
ow	oe	TI	3431	8	ain	aîΩ	MRE	0	0
ar	aΨ	MT	52	0	cch	1cΦ	PEH	13	0
or	oΨ	TT	191	0	c2ch	2cΦ	ſEH	1144	782
aw	ae	MI	1265	5	8ar	8aΨ	CIT	19	0
8h	8Φ	CH	3248	0	ain	aiΩ	MſE	2	0
four-glyph "words"					five-glyph "words"				
cchh	1c8Φ	PENH	0	0	tlain	ΔhaiΩ	ETIſE	0	0
c2ch8	2c8Φ	ſENH	0	0	tlch8	Δhc8Φ	ETENH	0	0
8ain	8aîΩ	CIRE	0	0	tlain	ΔhaîΩ	ETIRE	0	0
tlcch	Δhĉ Φ	ETHH	0	0	tlcch8	Δhĉ8Φ	ETHNH	0	0
tlaw	Δhae	ETII	215	52	olain	ohaîΩ	TTIRE	0	0

Source: author's analysis. Counts of Albanian strings based on Buzuku, Gjon: *Meshari* (1555).

So far, we have some evidence to support the idea that the Voynich manuscript is derived in some systematic way from documents in Old Albanian. We can only test this hypothesis by attempting some mappings. To this end, we selected the five most common "words" of three glyphs, four glyphs and five glyphs in Language B and mapped them into letter sequences in Old Albanian, on the "Gold Bug" approach and including "marker glyphs." To give the hypothesis a sporting chance, we then searched *Meshari* for these letter sequences, whether or not they occurred

as whole words (since we do not know whether "words" in Voynich are really words). The results were as shown in Table 13.2.

Many of the mapped text strings looked pronounceable, but most of them rarely occurred in *Meshari*, at least not in their precise letter sequences and not as words. Only "ſeh" and "etii" showed up frequently as words in *Meshari*; and we have to wonder if that was by chance (the same way that the English word {okay} shows up in the v101 transliteration). For what it is worth, the most-common words of these lengths in *Meshari* are as follows:

- three letters TEH, PER, MEH, ƐOT, SEH
- four letters ƧIΞE, ATAH, TINE, ATEH, ENDE
- five letters ENDEH, ANſTE, ENBEH, NUCHE, ENBII[18]

We can imagine that a little cribbing would produce some more robust transliterations. If, for example, ⊂ᑕc୨ [1cΦ] maps to TEH, MEH, SEH or ſEH, then ⊂ᑕc8୨ [1c8Φ] should map to a string like TE?H, ME?H, SE?H or ſE?H. Such words exist in *Meshari*; the most common are TECH (119 occurrences) and TEIH (seventeen occurrences). That suggests a crib:

initial ⊂ᑕ [1] ⇔ T

internal c [c] ⇔ E

final ୨ [φ] ⇔ H

internal 8 [8] ⇔ C.

We could continue the cribbing process among the longer glyph strings and words. This seems like a promising area for research.

GLYPH PAIRS TO DIGRAPHS

Alternatively, we can try (as we did in Chapter 8) a mapping of glyph pairs to digraphs. We therefore matched up the glyph pair frequencies with the digraph frequencies in *Meshari*, with the results summarized in Table 13.3.

These correlations were quite encouraging. In particular the correlation of final glyph pairs with final digraphs was better than that of the three languages that we had tried in Chapter 8 (medieval Italian, Latin, and Old French). This was largely due to the ubiquity of the final EH in Old Albanian, which mirrored the frequency of the final ꝺ9 [8Φ] in the Voynich manuscript.

To elaborate these correlations into transliterations, as before our best hope was to look at Voynich "words" of four glyphs, which should parse into first pair + second pair. For the three most common four-glyph "words," this yielded the following results:

Table 13.3. Voynich manuscript, Language B, v104 transliteration: position-dependent counts of the most common glyph pairs, and of the most common digraphs in Old Albanian

Voynich Language B				*Meshari*		Voynich Language B				*Mesh*
glyphs	v101	v104	count	digraph	count	glyphs	v101	v104	count	digraph
isolated position (two-letter words)						initial position				
ol	oe	oe	555	EN	1,337	ᖘolᖴ	4oh	Δh	2,449	TE
aᖍ	ay	aΨ	404	ƬI	1,277	ctc	1c	1c	1,950	EN
oᖍ	oy	oΨ	309	HI	983	oᖴ	ok	ok	1,444	PE
al	ae	ae	297	EM	184	olᖴ	oh	oh	1,397	ME
ꝺ9	89	8Φ	195	TE	135	c̊tc	2c	2c	1,335	CH
ᖘol	4oe	Δe	158	ME	105	ꝺa	8a	8a	1,119	ƐO
ct9	19	1Φ	50	Tɣ	85	ᖘolᖴ	4ok	Δk	825	ſE
cto	1o	1o	38	NE	31	ol	oe	oe	729	TI
c̊t9	29	2Φ	36	IN	20	cto	1o	1o	410	TU
ᖘolᖴ	4oh	Δh	30	DE	19	ɬlᖴ	eh	eh	398	AT
				R squared	85.3%					R squared
internal position						final position				
c8	c8	c8	3,897	NE	4,503	ꝺ9	89	8Φ	5,373	EH

⁋a	ha	ha	2,142	TE	4,177	⟡	m	îΩ	2,518	NE
⁋c	hc	hc	1,478	ET	3,910	c9	c9	cΦ	1,828	TE
cc	1c	1c	1,434	EN	3,547	a2	ay	aΨ	1,682	IH
cc	cc	cc	1,152	ER	2,991	⟡	n	iΩ	1,431	ER
⁋cc	hC	hĉ	1,059	IT	2,516	a4	ae	ae	1,248	AH
⁋a	ka	ka	1,054	UN	2,506	o4	oe	oe	1,084	IT
co	co	co	1,024	ND	2,379	cc9	C9	ĉΦ	875	ET
8a	8a	8a	1,019	IE	2,140	o2	oy	oΨ	546	II
⁋c	kc	kc	802	UO	1,853	8	p	iΠ	517	UE
				R squared	86.8%					R squared

Source: author's analysis.

Note: The counts of digraphs are based on Buzuku, Gjon: *Meshari* (1555). This table shows only the ten most frequent glyph pairs and digraphs in each position. R squared refers to the correlations based on the sixty-two most frequent glyph pairs and digraphs in each position.

> cc8 [1c8Φ] ⇔ cc [1c] + 8 [8Φ] ⇔ ENEH in Old Albanian, which is a real word, occurring twenty-two times in *Meshari*, and, with thanks to Dr. Michiel de Vaan of the University of Lausanne, could be a form of the possessive "our"[19]
>
> ĉc8 [2c8Φ] ⇔ ĉc [2c] + 8 [8Φ] ⇔ CHEH in Old Albanian, a real word but rare, occurring only six times in *Meshari* and possibly meaning "you have"[20]
>
> 8a ⟡ [8aîΩ] ⇔ 8a [8a] + ⟡ [îΩ] ⇔ ƐONE in Old Albanian, a real word occurring 123 times in *Meshari* and meaning "Lord" in the accusative case[21]

If we have really unearthed a religious interpretation for the ubiquitous glyph string 8a ⟡ [8aîΩ], this would be a significant advance toward an understanding of the meaning and purpose of the Voynich manuscript.

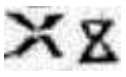
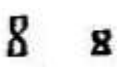
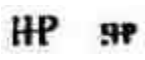
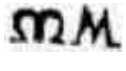
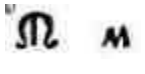
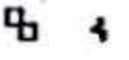
With that, we turn to the eyeball approach. Do any of the Hieronymus letters resemble any of the glyphs in the Voynich manuscript?

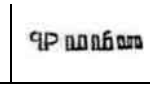
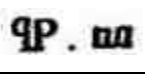
THE EYEBALL APPROACH

Here we have no mathematics or statistics to help us out. We are dependent on a subjective perception of what the Voynich glyphs and the Hieronymus letters look like. Where one person may see a resemblance, another may not.

We set out therefore to look for the kind of evolution that we imagine might have occurred over centuries, wherein the squarish and blocky script of Hieronymus might have been transformed into the rather elegant and rounded glyphs of the Voynich manuscript. Perhaps an analytical approach could be devised to this end, using the technologies of optical character recognition and reverse image search. Like many other tasks that we have considered, that is beyond the scope of this book. At this point it has to be said that the search for resemblances was a work of imagination rather than of analysis.

Table 13.4. Possible visual resemblances between some common Voynich glyphs and selected letters in the Hieronymus script

Voynich glyph			Hieronymus alphabet					
glyph	v101	v104	Postel	Rocha	Duret	Diderot & D'Alembert	Fournier	pronunciation
			1538	1591	1613	1751	1766	
ﻉ	8	8						I
ﻫ	h	h						Y
ﻣ	m	îΩ						M
ﻁ	4o	Δ						K

ff	k	k						V

Sources: author's analysis; Hieronymus upper-case and lower-case characters as indicated.

Table 13.4. Possible visual resemblances between common glyphs in the Voynich manuscript, and letters in the Hieronymus script of the Illyrian language. *Source: v101 alphabet from Glen Claston; v104 alphabet by author; Hieronymus characters from sources indicated in Figure 13.6*

We had to keep in mind that in cases where we saw resemblances, we had to map relatively common glyphs to relatively common letters. In all Indo-European languages, the vowels A, E, I, O (and sometimes U) are among the most-common letters. Therefore, the Hieronymus characters corresponding to these vowels must map to relatively common glyphs.

In this author's view, there are only five visual resemblances between common Voynich glyphs and letters in the Hieronymus script, as summarized in Figure 13.4.

If the Voynich scribes worked from documents in the Hieronymus script, those documents must predate the Voynich manuscript. As we observed earlier, as late as 1766 the Hieronymus characters had a blocky, archaic look to them, as if they were designed to be inscribed in stone with a chisel, and not written with a quill or pen. The Voynich glyphs look as if they have a history of refinement, to arrive at the elegant shapes that we have seen. If this observation has any merit, we will have to wonder whether the Voynich manuscript, though undoubtedly written on parchment from the fifteenth century, acquired its text after the Fournier document was published: that is, after 1766. We can make of that what we will.

If it is within the realms of possibility that the Voynich manuscript dates from any time later than 1637, then we would be obliged to infer that Kircher did not receive an extract from it. Perhaps the attachment to the original letter of Moretus was not, as we had assumed, a copy from the Voynich manuscript, but was from some entirely unrelated document, written in the Illyrian language (or an Illyrian language) and in the Hieronymus script. If so, Kircher might not have considered it a priority, and (as he alluded in his reply to Moretus) might never have found the time

or interest to translate it. Unless further correspondence is unearthed one day from Bohemian or Roman archives, we shall never know.

In our epilogue we shall set out a series of methodologies whereby interested readers could search for the meaning of the Voynich manuscript, or for any parts of it that appeal to them, according to their predilections. They can work on any of the underlying languages that have been mentioned in this book, and indeed they can try others. They can use the v101 or other alphabets that we have proposed, or they can invent new ones: the glyphs certainly allow differences of interpretation, even those as fundamental as what constitutes a glyph. Indeed, they can dispense with all the computer-readable transliterations and work with the original glyphs, using the Voynich v101 Unicode set: a task which will repay the effort.

Perhaps then one of our readers will break the bars of the manuscript, a feat that Georgius Barschius hoped that Athanasius Kircher would accomplish, as he implored in his letter almost four hundred years ago.

CHAPTER 14

VOYNICH 2022

After Mary D'Imperio's seminal conference on the Voynich manuscript in 1976, decades passed in which there was the barest minimum of scholarly study of the manuscript. Such papers as appeared to be both serious and accessible, this author attempted to summarize in Chapter 7, "Markov." Why scholars should have avoided investigation of the Voynich manuscript, we cannot say; except that in 2019, Dr Lisa Fagin Davis, of the Medieval Academy of America, characterized the academic perspective on the manuscript as follows:

> The manuscript is so infamous that medievalists and other scholars have been conditioned to roll their eyes when the very name is mentioned. It is easy to forget that underneath the media buzz and unsubstantiated theories lies an actual medieval object well worthy of study, six hundred years old, with a lengthy and fascinating recorded history.[1]

It was therefore something of a paradigm shift when, in the summer of 2022, the University of Malta invited scholarly papers on the Voynich manuscript, with a view to an online conference. That conference duly took place, with a nominal base in Qrendi, Malta, on November 30 and December 1, 2022. At the time of writing, this is the only conference on the Voynich manuscript that has taken place under the sponsorship of an academic institution. Two keynote addresses and sixteen research papers were presented. There were twenty-three authors, of whom ten were independent researchers and thirteen were affiliated with universities in Australia, Germany, Malta, and the United States.

This author made a provisional classification of the conference papers, as follows:

- plaintext or encipherment (four papers): examining whether or not the Voynich text is enciphered from a plaintext, or is itself a plaintext, albeit in an unknown script
- meaning or gibberish (four papers): on whether or not the Voynich text contains meaning, or is random nonsense (or equivalently, but with an implication of intent, a hoax)
- prose or poetry (two papers): the statistical and spatial properties of words and of certain individual glyphs, notably the "gallows" glyphs
- transliteration (two papers): proposals for new transliteration alphabets
- scribes and scripts (three papers): on the identifiable individual scribes and their script conventions
- ownership (two papers): on the ownership of the manuscript between Voynich's purchase in 1912 and Kraus's donation to Yale University in 1969; and on whether it was once owned by Emperor Rudolf II
- cultural affiliations (one paper): the tent or canopy as a motif in the manuscript.

In this chapter, we have attempted to summarize and to comment on some of the papers which were presented at Voynich 2022. We have

selected those which seemed to us productive in terms of opening avenues for further research.

A. PLAINTEXT OR ENCIPHERMENT

The reader will have observed that in this book, we have not dwelt on the idea that the Voynich manuscript might be enciphered (that is, transformed from a meaningful plaintext in a natural language into a ciphertext in which the meaning is still present but concealed). Wilfrid Voynich himself had no doubt that the manuscript was enciphered; he referred to it as the "Roger Bacon cipher manuscript," and to this day the Beinecke Library uses the term "Cipher Manuscript" as the catalog entry.

It seems to this author that any research in this direction is doomed to fail, or at the most to advance by miraculous guesswork. Our reasoning is simply that most of the glyphs in the Voynich manuscript have never been seen in any other document.

Let us take a real-life example of presumed encipherment. On December 1, 1948, an unidentified man was found dead on a beach in South Australia. He became known as the "Somerton Man." His story had two elements of mystery. The first was the identity of the man. In July 2022, the leading researcher on the case published what he believed to be the man's name; we can reasonably accept that this part of the mystery is solved.

The second part of the mystery related to a tightly rolled scrap of paper in the dead man's pocket, which bore the printed words "tamam shud": meaning, in the Farsi or Persian language, "it is finished." The paper had been torn from the last page of an English-language translation of the Persian poems known as the *Rubaiyat of Omar Khayyam*. Extraordinarily, the book from which the paper had been torn was found; one of its pages contained indentations of handwritten text, which under ultraviolet light were transcribed as follows:

WRGOABABD {the first letter might be M}
MLIAOI {either underlined or crossed out}
WTBIMPANETP {the first letter might be M}

{what looks like an X with a double strikethrough}
MLIABO AIAQC
ITTMTSAMSTGAB {followed by an underline}.

There is no dispute that the letters are from the Latin alphabet, which today is used as an official script in 143 states plus the European Union. It would be possible to search for the above strings of text in every language which uses the Latin script; perhaps someone has already done it. In one language or another, some short strings might make sense: for example, GOABA (Malagasy for "giant"), PANET (Catalan for "bun") and SAMST (Luxembourgish for "the same"). Otherwise, we can readily believe that the text, as a whole, makes no sense in any language. It has therefore become a part of the narrative of the Somerton Man that the text is enciphered in some way. What it means, to this day no one knows.

When we turn to the Voynich manuscript, we cannot recognize the script. Therefore, we have no basis for making any statement to the effect that the text is enciphered. Nevertheless, several papers presented at the Voynich 2022 conference invite us to think about encipherment. We may see where that takes us.

ENCIPHERED … OR NOT

Claire Bowern and Daniel Gaskell, of Yale University, presented a paper with the provocative title "Enciphered after all?"[2] The authors selected twenty-two methods of "text manipulation": which we can regard as equivalent to encipherment. Some illustrative examples are as follows, in which four of the authors' algorithms are applied to the first line of Shelley's "Ozymandias":

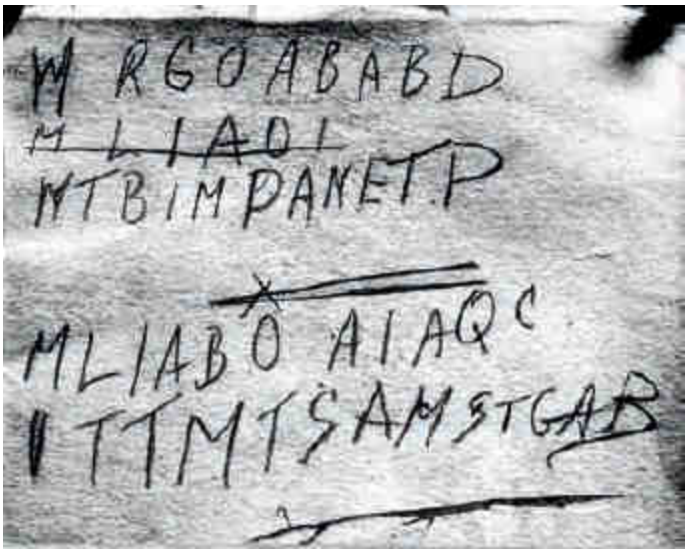


Figure 14.1. An enhancement of the indentations found on a copy of the *Rubaiyat of Omar Khayyam*, linked with the Somerton Man. *Public domain*

method of encipherment	*result*
none (plaintext)	I met a traveler from an antique land
adding the letter b after every letter	Ib mbebtb ab tbrbabvbeblbebrb fbrbobmb abnb abnbtbibqbubeb lbabnbdb
conflating the five most common consonants to “r”	I mer a rraverer from an arrique rard
removing all vowels	mt trvlr frm n ntqu lnd
in each word, sorting the letters in alphabetical order	I emt a aeelrrtv fmor an aeinqtu adln

Bowern and Gaskell applied each of these encipherments to a selection of twenty-five documents. All were in the Latin script; the languages included Albanian, Basque, Hungarian, Icelandic, Latin, Latvian, Slovak, Turkish, and some non-European languages. This resulted in a matrix of 22 ciphers × 25 documents = 550 document-cipher combinations. For each document and cipher, the authors calculated various statistical attributes, which they then compared to several versions of the Voynich manuscript. We might expect that if any document-cipher combination was found to be statistically similar to the Voynich manuscript, it would suggest both a source language and a cipher.

In this author's understanding of the paper, the results did not yield a match. The Voynich manuscript, in all its versions, was an outlier versus all the document-cipher combinations that were tested. The authors concluded (somewhat at variance with the title of their paper): "These text manipulations did not produce a single set of changes that mimic Voynich text."

AVE MARIA

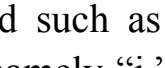
Jürgen Hermes, of the University of Cologne, authored a paper titled in part "The cipher that pretends to be an artificial language."[3] Hermes drew on what he described as the first printed book on cryptology, namely *Polygraphia*, written by the German priest and historian Johann Heidenberg (Johannes Trithemius), and published in 1518. Among the innovations in *Polygraphia* was the Ave Maria cipher, in which single letters were replaced by words. Hermes gave the example of encipherment of the English word "secret" by the string of meaningless words "pasil pasu pasi pasel pasu pasol." He then posed the question of whether a cipher like the Ave Maria could produce a text with similar statistical properties to those of the Voynich manuscript.

If this author understood Hermes's experiment correctly: he selected an arbitrary text (which was not identified) and enciphered it with two versions of the Ave Maria cipher. He then made statistical comparisons of the enciphered text with four documents in European languages: *The Mathematicall Praeface to Elements of Geometrie of Euclid of Megara* in Old English by John Dee, published in 1570; *La Divina Commedia* in Old Italian, by Dante Alighieri, completed about 1321; *Das Buch Paragranum* in German, by Theophrastus von Hohenheim (Paracelsus), published in 1565; and *Commentarii De Bello Gallico* in classical Latin, by Julius Caesar, published between 58 and 49 BC; as well as the sections of the Voynich manuscript in Language A and Language B.

Hermes concluded that "a cipher that replaces individual letters with invented words that are very similar to each other has the potential—in contrast to natural languages—to approximate the statistical properties of Voynichese. What this study still lacks is the simulation of a comprehensible selection process that can mimic a medieval encryptor."

We may interpret this finding as follows. If the Voynich manuscript was enciphered by some variant of the Ave Maria cipher, then individual Voynich "words" should represent individual letters in the source language. Now if the correspondence is one to one, then the number of distinct Voynich "words" should be the number of letters in the source language. The Voynich manuscript has about 8,100 unique "words"—the number cannot be exact and depends on which transliteration is used. But no known phonetic language, extant or extinct, has an alphabet of anything like 8,100 letters. The only languages with thousands of distinct glyphs are symbolic languages such as Chinese, in which the glyphs represent words.

That could constitute an interesting research task for a Voynich researcher who is familiar with written Chinese.

Alternatively, we could conjecture that nearly all the Voynich words are junk and that only the most frequent ones have meaning. In this case, and for the sake of argument, taking Latin as the source language, an ubiquitous Voynich word such as  {8am} could represent the most frequent Latin letter, namely "i." It might be instructive to take a chunk of Voynich text (necessarily a large chunk), code the twenty-three most common words to the twenty-three letters of the classical Latin alphabet (matching frequencies as far as possible), and discard all the other words. If something meaningful emerged, we would be onto something. If not, the same exercise could be repeated on other languages.

A KNOWLEDGE BOOK, A HOAX, OR A DISGUISE

Independent researchers Elonka Dunin and Klaus Schmeh adopted an entirely intuitive approach.[4] They assembled a database of 117 books known, or assumed, to have been written partly or entirely in cipher. The dates of writing or publication of these books spanned the fifteenth to twenty-first centuries. The authors classified these books by their apparent purpose or objective; by their date of creation; by their method of production (handwriting or printing); by whether they were known to have been deciphered, or not; by the method of encipherment, if known; and by

the number of authors, if known. They made the same classification of the Voynich manuscript.

By a process of intuitive comparison with the enciphered documents, Dunin and Schmeh formed the view that the Voynich manuscript was not a diary; not a treatise of a secret society; not a notebook; not a religious book; not a work of literature (such as a novel, memoir, or poetry); not an explicit puzzle or challenge for readers; and not an artwork. By elimination, they concluded that the manuscript was either a book of knowledge, or a hoax.

The authors added one alternative: "… the Voynich Manuscript could be a work disguised as a knowledge book about plants, astrology, and bathing, while the encrypted text covers something completely different. If this is the case, it would become much more likely that the Voynich Manuscript is a religious book or a piece of literature."

In earlier chapters of this book, we reported that in the herbal section, we could not identify a single name of a plant; nor in the "zodiac" section, a single name of a star. In the sections conventionally called "cosmological" and "pharmaceutical," we do not know what to look for in the text. So, there may be merit in the idea that the text is unrelated to the illustrations. It does not necessarily follow that the text is enciphered; it may simply be a plaintext in a script and a language that we have yet to identify.

B. MEANING OR GIBBERISH

The idea that the text of the Voynich manuscript is random nonsense or gibberish has been around for many years. Researchers and scholars have failed completely to generate meaningful text from any part of the manuscript longer than a few glyphs, except by resorting to arbitrary mechanisms. In this situation, it is natural to suspect that the text is without meaning. In the Voynich 2022 conference, seven authors in four papers explored this idea, and its converse: that the text has meaning.

VOYNICH AND ANGELIC LANGUAGE

Alexander Boxer, an independent researcher, presented his paper[5] on a comparison between the Voynich manuscript and a presumed example of gibberish: the writings in "Enochian" by the English mathematician Dr John Dee (the same John Dee who has appeared elsewhere in the story of the Voynich manuscript). Between 1581 and 1583, Dee transcribed an obscure text, apparently in whole or in part from dictation by his associate Edward Kelley. Dee believed (or wished others to believe) that he or Kelley was capable of receiving messages from angels, in the language of angels.

Dee's document is now known as *Sloane Manuscript 3188*[6] and is hosted by the British Library. The catalog entry for MS 3188 is titled *John Dee's conferences with angels*; the description reads in part as follows:

> 1r–169v: John Dee's conferences with angels, 22 December 1581–30 May 1583, in five books…. These are records of messages delivered by angels and other spirits, with the aid of a crystal ball…. Ashmole's preface explaining how the manuscript came into his possession. ff. 4r–108r: Dee's conferences with angels. ff. 109r–169v: Notes in cipher by William Shippen.

Dee transcribed the supposed communication with angels in a Latin script, which later came to be known as "Enochian." An example is reproduced in Figure 14.2.

The linguist Donald Laycock, in his book *The Complete Enochian Dictionary*, wrote: "All the facts seem compatible with Kelley pouring out a string of gibberish while in a trance state."[7] Let us then assume that it is gibberish, and let us follow Boxer's argument.

As a reference corpus, Boxer used Christopher Whitby's annotated edition[8] of Dee's work.

In addition, Boxer drew upon the following classical and medieval documents in European languages, all of which have undisputed meaning: the Torah, in Hebrew, probably completed in the fifth century BC; *Commentarii de Bello Gallico*, in classical Latin by Julius Caesar, published between 58 and 49 BC; *Der abenteuerliche Simplicissimus Teutsch*, in German by Hans Jakob Christoffel von Grimmelshausen, published in

1668; and *The Canterbury Tales*, in Middle English by Geoffrey Chaucer, written between 1387 and 1400.

By means of a series of statistical tests on structure and vocabulary, Boxer found that the Enochian text was greatly dissimilar, and the Voynich manuscript generally similar, to the meaningful documents. In other words, the Enochian text bore the fingerprint of gibberish, and the Voynich manuscript did not. A statistical chart in Boxer's paper makes this distinction very clearly.

Boxer concluded: "The Voynich {manuscript} does not exhibit this statistical fingerprint {of *hapax legomena*}, and by this metric appears more similar to natural language than to Enochian…. Our observation does not eliminate the possibility that the Voynich {manuscript} may belong to another class of gibberish which is substantively different from the Enochian of Sloane MS 3188."

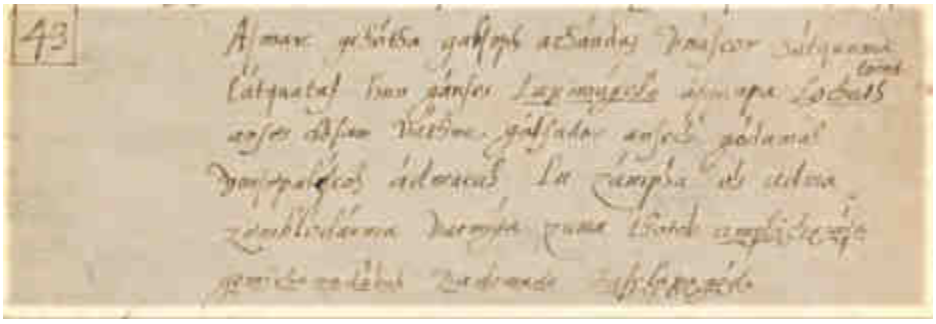


Figure 14.2. An extract from folio f79v of Sloane Manuscript 3188. *British Library*

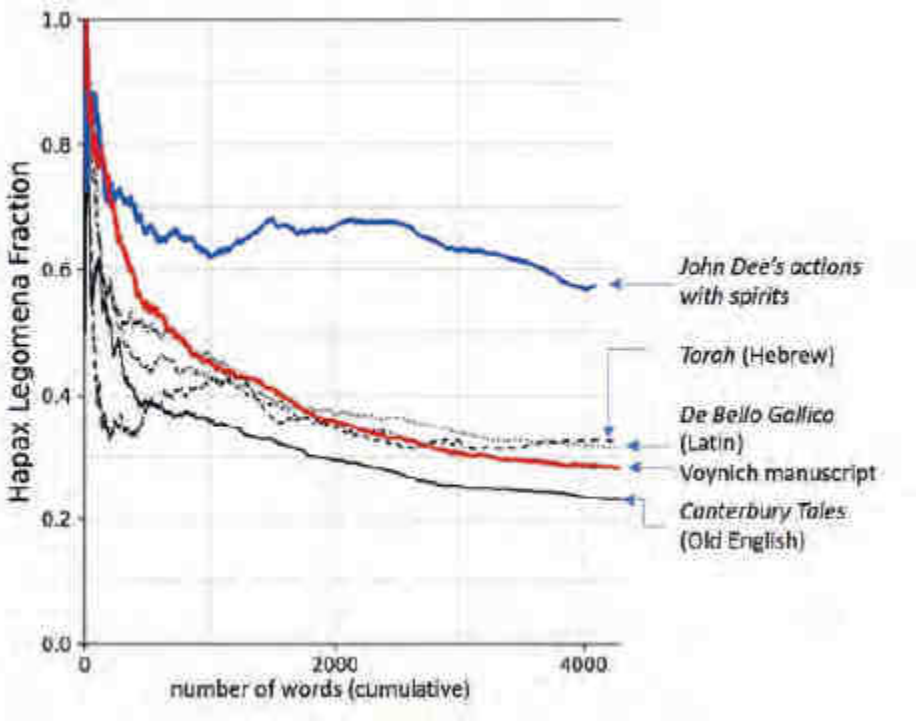


Figure 14.3. The running fraction of hapax legomena (words which occur only once) in the Enochian text, the Voynich manuscript, and selected European texts. *Alexander Boxer*

VOYNICH AND INTUITIVE GIBBERISH

Daniel Gaskell and Claire Bowern of Yale University conducted an intriguing practical experiment which they reported in their paper titled in part "Gibberish after all?"[9] They assembled forty-two volunteers, some of whom were familiar with the Voynich manuscript. They asked the volunteers to write short handwritten documents in lowercase Latin script, each of between one and three pages, in which the content intentionally would be gibberish. Each volunteer was permitted to define what he or she considered gibberish; but was invited to make the gibberish resemble a natural language. They were allowed to include illustrations. An extract from one of the results is shown in Figure 14.4.

Gaskell and Bowern then deployed a massive battery of statistical comparisons of the volunteers' gibberish documents with the Voynich manuscript (in several different transliterations); with sixty meaningful documents in historical and modern natural languages; and with fifteen documents in constructed languages such as Esperanto.

They found that statistically, human-produced intuitive gibberish was quite similar to natural languages; and that the Voynich manuscript was statistically similar to both to the intuitive gibberish and to natural languages on nearly all of the metrics.

They concluded: "… intuitively-generated gibberish can both adequately replicate many {statistical} features of natural language and produce a wide range of other unusual properties … If the VMS {Voynich manuscript} is indeed gibberish, … it is perhaps not surprising that it would be both language-like and statistically unusual."

WORD PAIRS IN THE VOYNICH MANUSCRIPT

Andrew Caruana, Colin Layfield, and John Abela, all from the Department of Computer Information Systems at the University of Malta, brought a computational approach to bear[10] on the frequencies of words in the Voynich manuscript. Specifically, they looked at the frequencies of word pairs.

In addition to the Voynich manuscript, the authors used as reference corpora the following documents: five books from the King James Bible, in Old English, published in 1611; *La Divina Commedia* by Dante Alighieri, in Old Italian, completed about 1321; and William Shakespeare's *Julius Caesar* (1599) and *Macbeth* (1623), both in Early Modern English. In the English and Italian documents, they removed numerals, punctuation and "stop words"—that is, common words such as, in English, "a," "an," "and," "the." In the Voynich manuscript, there are no recognizable numerals or punctuation, and the "stop words" have to be guessed by their frequency; we may suspect that the authors removed "words" such as {8am} and {am}.

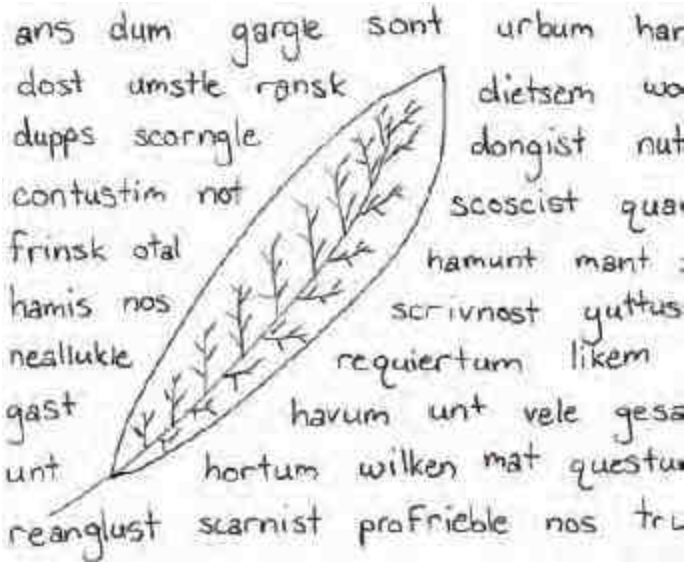

ans dum gargle sont urbum har
dost umstle ransk dietsem wo
dupps scorngle dongist nut
contustim not scoscist qua
frinsk otal hamunt mant
hamis nos scrivnost guttus
nealluke requiertum likem
gast havum unt vele gesa
unt hortum wilken mat questu
reanglust scarnist profrieble nos

Figure 14.4. An example of text intended to be gibberish. *Daniel Gaskell and Claire Bowern*

The authors used the concept of a skewed pair. This is a pair of words in which one is typically the first and the other the second. They gave the example of (brown, dog), which in English is much more frequent than (dog, brown). In each document, the authors calculated the frequency of skewed pairs and then shuffled the words in the document randomly and recalculated the frequencies. In a natural language, shuffling should break up the pairs and reduce the frequency of skewed pairs. This is what the authors found in their reference documents; but also in the Voynich manuscript.

The authors concluded: "This could indicate that the Voynich Manuscript is not random text but may be [written in] a language or a cipher." In other words, as in many previous statistical approaches, their tests were more consistent with the hypothesis that the Voynich manuscript contains meaning, than that it has no meaning.

WORDS AND THE SOURCE LANGUAGES

Luke Lindemann, of the George Washington University, presented a paper[11] with the specific objective of identifying the most-plausible language families that might underlie the Voynich manuscript. To do this, he

employed two statistical tests which focused on the words in a given manuscript.

The first metric was the Moving Average Type-Token Ratio (MATTR). In any given text, the type-token ratio (TTR) is the number of unique words divided by the total number of words. A few examples may illustrate this concept:

- The first line of Shelley's "Ozymandias" is "I met a traveler from an antique land"; it contains eight words, each of which is unique within the line; so, the TTR is 1.
- The first line of the famous soliloquy in *Macbeth* is "Tomorrow and tomorrow and tomorrow, creeps in this petty pace from day to day," with fourteen words and ten unique words; so, the TTR is 10/14 = 0.71.
- Abraham Lincoln's *Gettysburg Address* (in the most popular version) contains 268 words. In this version there are 141 unique words, from "a" to "years." So, the TTR is 141/268 = 0.53.

In general, the longer is the text, the lower is the TTR. The MATTR metric compensates for the length of the text by taking the average TTR over equal partitions of the text. Lindemann used a partition of 1,000 words.

The Most Common Words (MCW) metric is the proportion of the word count in a text that consists of the most-frequent unique words. The researcher can choose the number of words that he or she deems to be frequent; Lindemann set this number at 70. Generally, a language which simplifies its vocabulary has a lower MCW than a language with a more complex vocabulary. Lindemann gave the example of the English word "the," which accounts for 7 percent of all written words in English. In German, the equivalent "der" accounts for 4 percent of the written language, because "the" can also be "das," "die," and so on.

Lindemann applied these two metrics to the Voynich manuscript, separately for Language A and Language B; to 100 samples of text from each of ten modern language families (Celtic, Germanic, Indic, Iranian, Philippine, Romance, Semitic, Slavic, Turkic, and Uralic); and to samples

of text from four historical documents: *La Rettorica*, written around 1260 by Brunetto Latini in the Florentine vernacular (a Romance language), *Tanach Bereshit* in Hebrew (a Semitic language), probably from the fifth century BC, *Mabinogion* in Middle Welsh (a Celtic language) from between 1350 and 1410, and *Codex Suprasliensis* in Old Church Slavonic (a Slavic language) from the tenth century.

Lindemann's findings are well illustrated in one of the statistical charts from his paper, which is reproduced (with this author's annotations) in Figure 14.5. In summary:

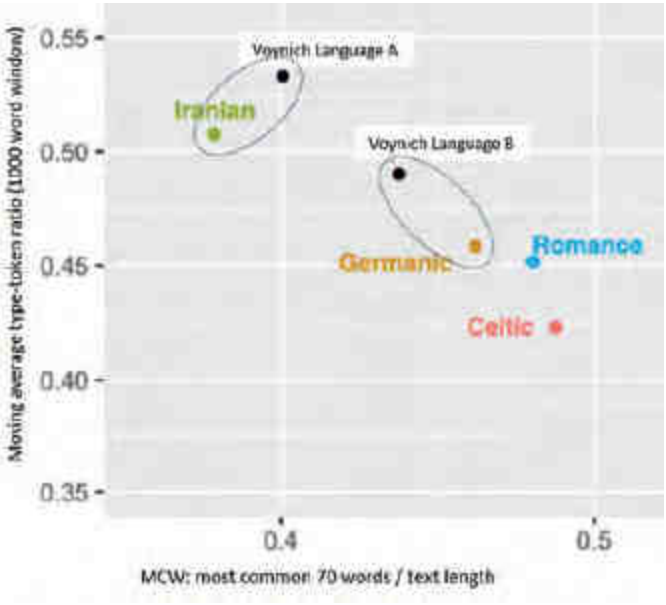


Figure 14.5. The positioning of the Voynich manuscript against modern language families, by the MATTR and MCW metrics. *Luke Lindemann, additional graphics by author*

- The sections of the Voynich manuscript in Language A were most closely positioned against the modern Iranian language family, which includes Kurdish, Pashto, Persian, and Tajik.
- The sections of the Voynich manuscript in Language B were most closely positioned against the modern Germanic language family, which includes Dutch, English, German, and the Scandinavian languages; the next closest modern family was the Romance group, including French, Italian, Spanish, and their relatives.

- Neither of the Voynich languages was closely positioned against any of the selected historical documents.

Lindemann concluded that "{the} Voynich {manuscript} does, in fact, encode meaningful text … improving upon this methodology may help to geographically situate Voynichese, either by identifying an underlying natural language or by suggesting the kind of language that a medieval hoaxer could have been imitating."

We could add that a possible Persian connection to the Voynich manuscript is not well explored. We recall that the manuscript was written on calfskin. In medieval times, most European cattle were descendants of animals domesticated in the Fertile Crescent (what is today parts of southwest Iran, central Iraq, and Syria). The calfskin of the Voynich manuscript has been tested by radiocarbon analysis, and its age thereby established within a few decades; if any DNA still exists, genetic analysis of the calfskin might yield an insight into its geographical origins.

C. PROSE OR POETRY

At least eight of the papers presented at the Voynich 2022 conference seemed to contribute to two long-standing strands of debate on the Voynich manuscript, namely plaintext vs. ciphertext, and meaning vs. nonsense. There might be merit in a third debate, arising from the "functionality of the line," which Prescott Currier first identified at the Voynich conference in 1976. Currier observed that the lines of the manuscript were, in some sense, stand-alone; a line did not see to wrap to the next line, as prose text normally does in natural languages.

Currier did not attempt to explore this functionality; and although most researchers of the manuscript seem to agree that the lines have some function, this author has not seen any concerted effort to determine what that function is. In natural languages, there is at least one type of text in which the line has a clearly defined function: and that is poetry (or equivalently, the lyrics of songs). We might propose that the third strand of debate on the Voynich manuscript should be: prose vs. poetry. In the

proceedings of the Voynich 2022 conference, there were two papers that seemed to contribute to such a debate.

THE STRANGENESS OF THE GALLOWS

Tavi Stafford is an independent researcher, who focused in her paper[12] on the strange properties of the tall and elegant symbols known as the "gallows" glyphs.

Stafford selected the following glyphs for study: {f}, {g}, {h}, {k}. She assembled samples of these four glyphs from the sections of the Voynich text written by Scribes 1, 2, and 3 as identified by Dr. Lisa Fagin Davis. For each sampled glyph, she recorded its characteristics, in the following respects:

- which scribe had written the glyph
- whether it was the first glyph of a "word," a line or a paragraph
- whether it was in the first row of a paragraph
- what were the immediately preceding and following glyphs

On the basis of these samples, she performed statistical tests from which she found a number of patterns or correlations (which she called "habits"), including the following:

- All three scribes had a preference for a gallows glyph as the first glyph of a paragraph.
- All four of the gallows glyphs, if not at the start of a paragraph, were rarely the first glyph of a "word."
- The glyph preceding a gallows glyph was very likely to be an {o}.

Stafford considered whether the gallows glyphs had undergone some kind of mechanistic transformation according to their position, and also whether they might simply be demarcations or ornamentation. She

concluded that there was no single process that could account for all the patterns that she had observed.

We could add one observation: what is missing in this analysis is a set of control documents. Viewed in isolation, the spatial patterns of the gallows glyphs might seem strange. But such patterns can be found in documents in natural languages: especially in highly structured documents such as poetry. For example, in Dante Alighieri's *La Divina Commedia*:[13]

- The three letters C, E, and P account for 38 percent of the first letters of the verses.
- The four letters O, E, I, and A account for 52 percent of the second letters of the verses.
- The letters A, E, O, and I account for 99 percent of the last letters of the lines (as we would expect for word endings in Italian).

In other words, the spatial patterns of the glyphs do not exclude the hypothesis that the Voynich text is in a natural language or languages.

THE STRANGE PLACING OF WORDS AND GLYPHS

Patrick Feaster, an independent researcher, made a new contribution[14] to the well-documented phenomenon that glyphs, glyph strings, and "words" seem to have preferences for certain positions in the Voynich manuscript.

Feaster introduced two metrics to measure the positions of graphemes (using this term to encompass a single glyph, a string of glyphs, or a "word"):

- rightwardness: meaning the position of the grapheme in the line, measured from the left, scaled so that the first grapheme in the line has the value 0 and the last grapheme has the value 1.
- downwardness: meaning the position of the grapheme in the paragraph, measured from the top, as a ratio of the number of

lines in the paragraph, scaled to the range 0 (occurring in the first line) to 1 (occurring in the last line).

Feaster used these metrics, separately or together, to characterize the spatial positioning of selected graphemes. He concluded: "Numerous glyph types can be shown to appear preferentially in certain parts of lines and paragraphs as opposed to others." As to what these patterns might mean, he suggested the following:

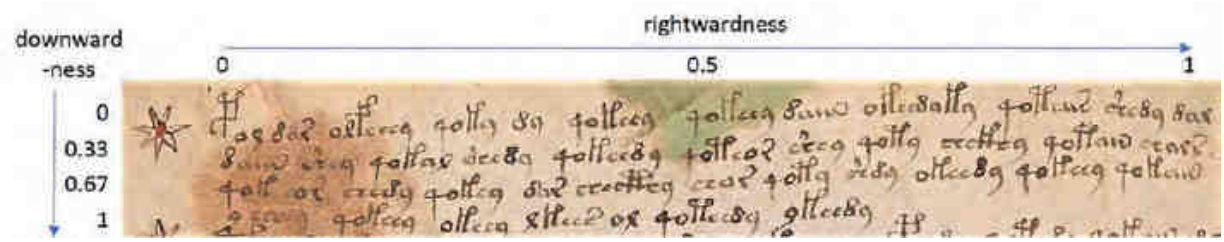


Figure 14.6. Voynich manuscript, folio f103v, paragraph 1. *Beinecke Rare Book and Manuscript Library, additional graphics by author*

One {explanation} is that each line of text corresponds to some unit of meaningfully patterned content, such as a grammatical sentence, a line of poetry, or an entry in a list.

He gave an example from *Aeneid*, an epic poem written in classical Latin by Publius Vergilius Maro (popularly known as Virgil) between 70 and 19 BC. In *Aeneid*, words ending in "-es" had a rightwardness score of 0.399, and words ending in "-ibus" had a score of 0.799—what Feaster called "a difference as stark as any presented above {from the Voynich manuscript}."

Feaster's findings are an impetus to further study of the Voynich manuscript as a possible work of poetry, in which poetic conventions drive the placement of words, and specifically their horizontal and vertical positions relative to paragraphs, pages and the work as a whole.

D. SCRIBES AND SCRIPTS

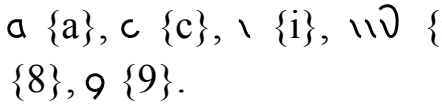
As we recall from Chapter 3, Dr. Lisa Fagin Davis' keynote address extended Prescott Currier's early work on the "hands" or "scribes," and completed Currier's work on the allocation of each of the Voynich pages to Language A or Language B. In addition, Katie Painter and Claire Bowern, of Yale University, attempted to place the writing patterns of the scribes within the known families of medieval scripts, on the basis of the shapes, sizes, and ornamentation of the glyphs.

THE EVOLUTION OF THE GLYPHS

Painter and Bowern presented a paper[15] on the application of phylogenetic methods to the Voynich glyphs. In biology, phylogenetics is the study of the evolution of the relationships between living organisms. The authors' idea was that like living creatures, scripts evolve over time, changing their shapes, styles, and relative positions; and that the Voynich glyphs might fit within an evolutionary pattern that could be recognized from other scripts.

To this end, the authors examined the five scripts employed respectively by the five Voynich scribes, as identified by Dr. Lisa Fagin Davis; and compared these with the scripts of fifty-two manuscripts written between the years 700 and 1500. Of these manuscripts, the majority were written entirely or partly in Latin; other languages represented were Arabic, Bohemian, English, French, German, Greek, Italian, and Polish. The script families were those listed in Figure 14.7 (with illustrative examples, which are drawn from the specific source documents referenced in the paper, but do not appear in the paper itself):

For comparison, Painter and Bowern selected ten Voynich glyphs which, in their judgment, most closely resembled Latin letters or Roman numerals, as follows:



The authors coded each of the European scripts and each of the selected Voynich glyphs in terms of selected attributes, relating to how the symbols were constructed. These attributes included the shape, the size, whether the

symbols had ligatures (joining them to preceding or following symbols), whether they extended above or below the line, and whether they had serifs (small strokes attached to the ends of larger strokes). They then conducted a range of statistical tests, broadly speaking to see whether the Voynich glyphs seemed to be related to a European family of scripts.

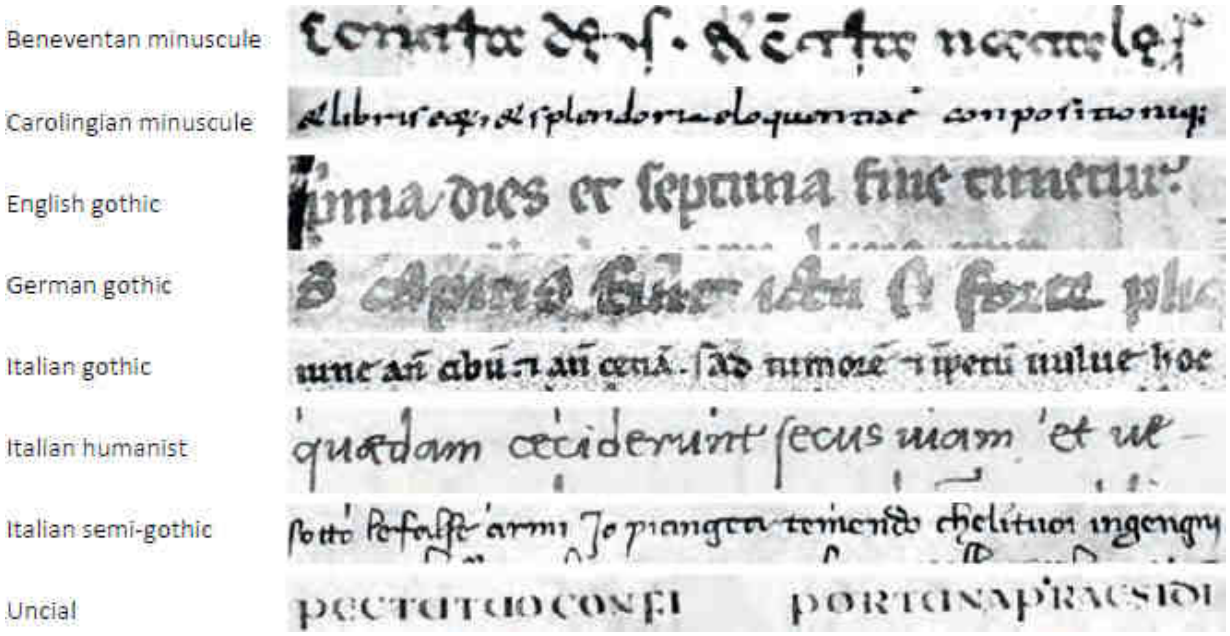


Figure 14.7. Examples of scripts used in selected documents referenced by Painter and Bowern (2022). *Yale University Library Digital Collections, additional graphics by author*

The authors found that in terms of shape, the selected Voynich glyphs were closest to the Beneventan family. In terms of size, the glyphs did not fit with any of the European scripts. By virtue of the rarity of ligatures, the glyphs most closely resembled the Uncial family. As a whole, the Voynich glyphs did not seem to belong to, or to be inspired by, any of the script families that were studied.

E. WHAT IS A GLYPH?

Probably the most fundamental paper in the Voynich 2022 conference was that presented by Massimiliano Zattera, an independent researcher. His

paper was titled “A New Transliteration Alphabet Brings New Evidence of Word Structure and Multiple “Languages” in the Voynich Manuscript.”[16]

Zattera’s paper went to the heart of an issue that, in this author’s view, had been left unresolved in Voynich research: namely, what is a glyph?

We can see that the text of the Voynich manuscript is made up of strings of symbols separated by spaces; since these strings look like words in natural languages, we generally call them “words.” The symbols in turn are made up of multiple strokes of the quill; and if we find a set of quill strokes that does not touch any adjacent set, we usually call it a “glyph.” We are tempted to think of a glyph as a letter, in the sense of a letter in a natural language, meaning a symbol that represents a sound. But we do not know whether that is the case, or not.

This dilemma is well illustrated by some of the differences between the various transliteration alphabets that have been developed over the years. The alphabets in widest use today are EVA (Extensible Voynich Alphabet) and Glen Claston’s v101. These two alphabets take a substantially different approach to interpretation of the glyphs, as the following examples may illustrate:

- This set of quill strokes occurs 4,063 times in the Voynich manuscript. Reasonable persons may disagree as to whether the quill strokes touch each other. In EVA, it is assumed that they do not, and therefore the set is transliterated as three letters: {iin}. In v101, it is assumed that this set of quill strokes is a single glyph: it is transliterated as {m}.
- This set of quill strokes occurs 906 times. Here, there is no question that some of the quill strokes are superimposed. In EVA, the set is assumed to consist of three letters: {c}, {t}, and {h}, and the whole is transliterated {cth}. The v101 alphabet treats this symbol as a single glyph, transliterated as {K}.
- This symbol is represented in EVA as {d}. But in v101, there are four versions of this symbol, with minor differences which might reasonably be interpreted as differences in handwriting: {8} occurring 10,363 times, {7} occurring 2,722 times, {6} occurring 182 times, and {&} occurring 35 times.

To this author's mind the essential differences between EVA and v101 are those illustrated above: that in some cases EVA has strings of letters where v101 has one, and that in some cases v101 has multiple letters where EVA has one.

Differences such as these, if left unresolved, will play havoc with any attempt to define words, to measure word frequencies, and make comparisons between the Voynich manuscript and natural languages, and even between the Voynich and gibberish.

Zattera's paper proposed to create a new transliteration alphabet. He started with EVA; and although he did not mention v101, his approach followed Glen Claston's, in that he simplified many of EVA's two-letter and three-letter concatenations into single glyphs. The result was very similar to v101.

Where Zattera went beyond previous transliteration alphabets is that he attempted to establish something like an alphabetical order. This is not an alphabetical order in the everyday sense, as used in languages in phonetic scripts, exemplified by "a, b, c, d, e, f," and so on. Rather, it is an order of precedence for glyphs within words.

Many Voynich researchers have observed that certain glyphs predominate as the initial and final letters of words, and that certain glyphs are typically preceded or followed by certain other glyphs. One of the most egregious examples is the glyph ϯ {4}, which in 97.4 percent of its occurrences is followed by the glyph o {o}.

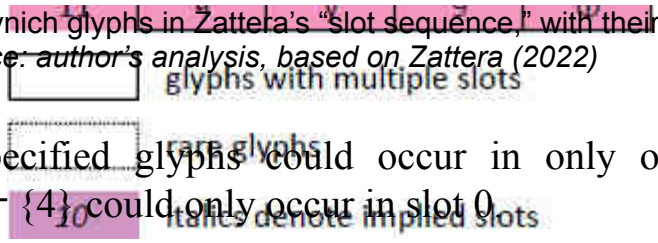
By computational processes which we need not summarize, Zattera constructed a "slot alphabet" consisting of twenty-six distinct glyphs, and a "slot sequence" consisting of twelve slots. He labeled the slots arbitrarily from 0 (the leftmost slot) to 11 (the rightmost slot). He asserted that every Voynich word could be constructed from these slots. A slot might contain a glyph, or it might be empty; but if it contained a glyph, it could only contain one of several specified glyphs. Figure 14.8 is a graphic which illustrates the "slot sequence."

In essence, the "slot sequence" stated that in any Voynich word, the following rules applied:

glyph slots as defined by Zattera alphabet

slot	glyph	EVA	v101	v104
0		q	4	Δ
0		d	8	8
0		s	s	s
1		y	9	Φ
1		o	o	o
2		l	e	e
2		r	y	y
3		f	f	f
3		p	g	g
3		k	h	h
3		t	k	k
4		ch	1	1
4		sh	2	1'
5		cfh	F	f1
5		cph	G	g1
5		ckh	H	h1
5		cth	K	k1
6		e	c	c
6		ee	C	ĉ
6		eee	d	d
7		d	8	8
7		s	s	s
8		a	a	a
8		o	o	o
9		i	i	i
9		ii	I	î
9		iii	il	ii
10		d	8	8
10		l	e	e
10		n	N	Ω
10		in	n	iΩ
10		iin	m	îΩ
10		m	p	Π
10		r	y	Ψ
11				

Figure 14.8. The Voynich glyphs in Zattera's "slot sequence," with their transliterations in EVA and v101. *Source: author's analysis, based on Zattera (2022)*



- Twenty specified glyphs could occur in only one slot: for example, ϯ {4} could only occur in slot 0.
- Six specified glyphs could occur in more than one slot: for example, ȣ {8} could occur in slots 0, 7, and 10.
- A given glyph could be followed only by a glyph in a higher slot.
- A given glyph could only be preceded by a glyph in a lower slot.

Zattera counted 31,317 words in the Voynich manuscript, excluding words which he considered rare, unreadable, or subject to differing interpretations. He asserted that the conformity of the Voynich words with the "slot alphabet" and "slot sequence" was as follows:

- 27,114 words, or 87 percent of the total, conformed precisely
- 3,249 words, or 10 percent of the total, could be separated into two parts each of which conformed
- just 954 words, or 3 percent of the total, did not conform.

Although Zattera did not present the frequencies of the individual glyphs in his "slot alphabet," it is straightforward to do so. Figure 14.9 is a graphic with the approximate counts of each glyph in the v101 transliteration, grouped according to their positions in the "slot" sequence.

This graphic illustrates that we could expect to construct common Voynich words from the common glyphs, in the order shown (but not in any other order). For example, ⊂⊂ȼȣ {1c89}, which corresponds to slots 4-6-7-11, should be a common word, and indeed it is, occurring 363 times. However, the sequence ȼ⊂⊂ȣ {c189}, with its glyphs in slots 6-4-7-11 (the "wrong order"), never occurs as a word.

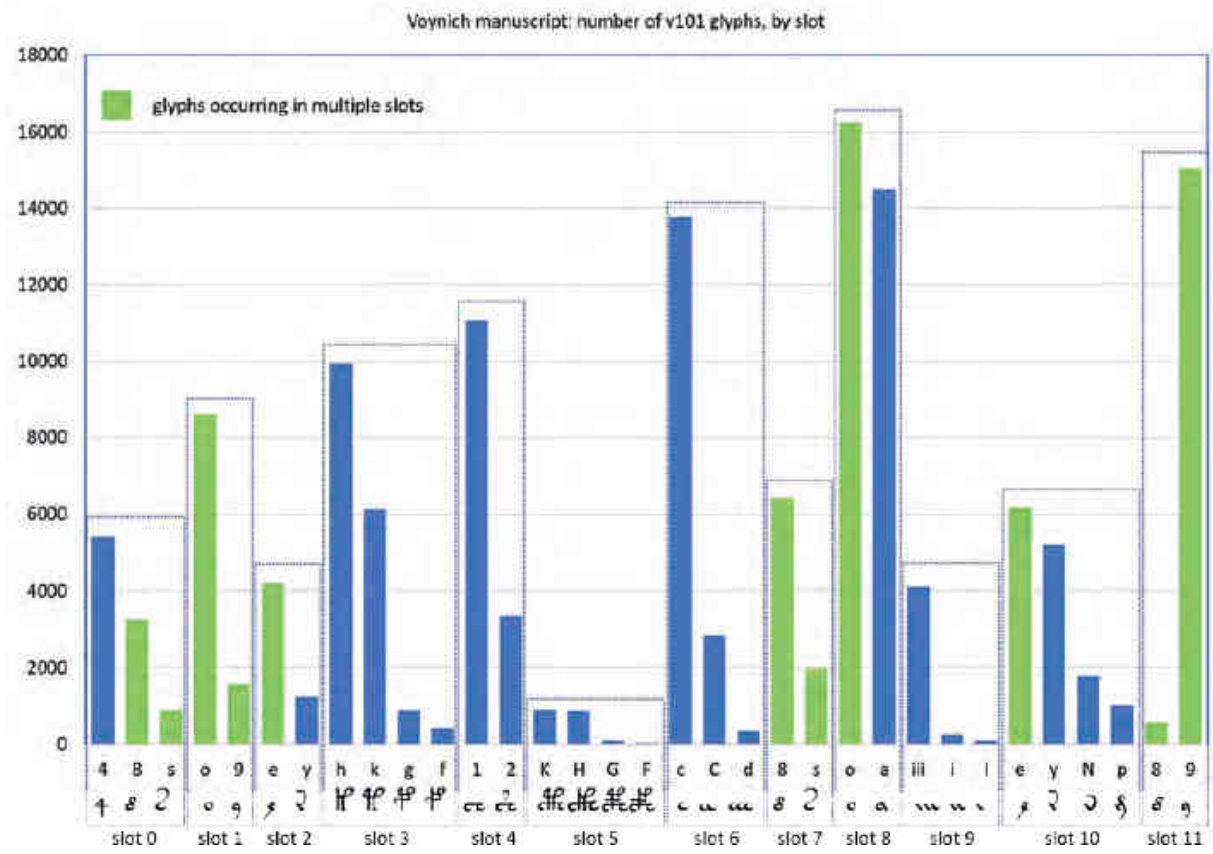


Figure 14.9. Numbers of glyphs in the Voynich manuscript, arranged by position in Zattera's "slot sequence." *Source: author's analysis based on v101 transliteration and Zattera (2022)*

This raises the intriguing possibility that in each Voynich word, the glyphs have been reordered from their original sequence. From another paper presented at the Voynich 2022 conference, we may recall the "sort_alph" cipher, in which each word was reordered in the alphabetical order of the letters. As an illustration, we applied this cipher to the first line of Shelley's "Ozymandias":

I met a traveler from an antique land

with the following result:

I emt a aeelrrtv fmor an aeinqtu adln.

Readers who are familiar with crossword puzzles or anagrams would not find it difficult to reconstruct the original text.

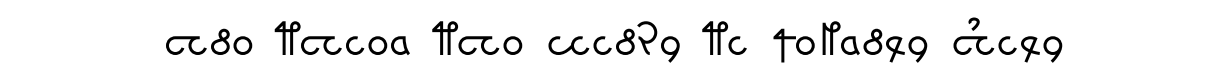
Something similar, but more complex, could be at work in the Voynich manuscript. We could imagine a Voynich scribe starting with a plaintext in a natural language (let us say, Italian as written in the fifteenth century). On his desk (if we assume that the scribe is male), or on a wall, he has a chart which maps Italian letters one to one (as we hope) to Voynich glyphs. Taking one Italian word at a time, the scribe duly replaces each letter with a glyph. Then he refers to another chart which defines the "slots," or the order of precedence of the glyphs. He reorders the glyphs. He writes the first "slot 0" glyph, if there is one; if there is none, then the first "slot 1" glyph, and so on, until the whole string of glyphs is reordered. Then he writes the fair copy on the manuscript.

If there is more than one glyph in a given slot, the scribe must have a rule for reordering within the slot. For example, let us suppose that the Italian word yields an o {o} and a 9 {9}. Both of these glyphs can be in slot 1; o {o} can also be in slot 9; 9 {9} can be in slot 11. The rule might be to write o9 {o9} so that o {o} precedes 9 {9}, as the broader slot sequence requires.

As an entirely hypothetical exercise, without any pretense to a scientific approach, we may turn to Dante Alighieri's *La Divina Commedia*, which was completed about 1321, and we may suppose, would have been widely known in central Italy in the early fifteenth century. We may imagine a scribe of the Voynich manuscript applying the "slot" treatment to the first line:

Nel mezo delcamin dinrã uita.

Having regard to the frequencies of the most-common letters in the medieval Italian language, we may make a guess at which letters would match up with which Voynich glyphs. Having done that, we reorder the glyphs in the order of their slots. There are of course multiple possibilities. One of this author's experiments yielded the following sequence:



Somewhat unexpectedly, these strings bear some visual resemblance to Voynich text. Two of these strings (the first and the third) are real Voynich words, though rare ones; the first "word" occurs just once, and the third "word" three times. Are we onto something?

TAKE-AWAYS

We may now make some general observations on the findings presented at the Voynich 2022 conference, as follows:

- The papers on the presence or absence of meaning generally favor the hypothesis that, in whole or in part, the text of the Voynich manuscript possesses meaning.
- Several papers grapple with the, so far unexplained, spatial structure of the manuscript: in particular, the tendency of glyphs, glyph strings, or words to appear in certain positions relative to lines and paragraphs. This suggests a case for a concerted effort of comparison of the Voynich manuscript with known historical works of poetry or song.

Zattera's paper on the "slot sequence" raises the distinct possibility that in each of the Voynich "words," as they appear in the manuscript, the glyphs have been reordered in something like an alphabetical order. We could think of this reordering as a very simple cipher. If this is the case, we have no way of re-creating the original order. We could adopt a trial-and-error method in which we start with selected plaintexts in natural languages; convert them to Voynich glyphs on the basis of matching letter frequencies with glyph frequencies; reorder the glyphs according to the "slot sequence"; and see whether any real Voynich "words" appear. It might be the only way.

EPILOGUE

HOWEVER IMPROBABLE

Sherlock Holmes liked to remind his long-suffering friend Dr. Watson that in any mystery, when one had eliminated the impossible then whatever remained, however improbable, must be the truth.

In this book, we have only attempted to eliminate the more improbable, and thereby to focus on the less improbable.

Among the narratives that surround the Voynich manuscript, our research and analysis as recounted in the preceding chapters have led us to believe that the following are among the more improbable:

- that the text is devoid of any narrative meaning
- that the text is enciphered from a plaintext in a natural language
- that the text is a work of continuous prose
- that the illustrations and the accompanying text are necessarily related to each other

We have therefore arrived at the following narratives that we think are less improbable:

- that the text, or some substantial part of it, has meaning
- that the text is a plaintext, either written in an embellished and abbreviated natural language or transliterated by some process from a preexisting script in a natural language
- that the underlying languages are European, the least improbable being those spoken and written in fifteenth-century Italy, or closely related languages
- that the document is a work with a poetic and possibly a rhyming structure, in which the illustrations are no more than allusions to the text, or possibly are unrelated to the text

It is always conceivable that the manuscript has a meaning, but not in a natural language. One can imagine a meaning in, for example, a mathematical language. Conceivably the 158,947 glyphs of the Voynich manuscript represent the first 158,947 digits of the decimal expression of the irrational number pi. We hope not.

Perhaps at this point some of our readers have seen a glimmer of light at the end of the Voynich tunnel. Whoever was the author, he or she had a sense of the structure, rhythm, and resonance of a natural language; and the scribes who put this work onto parchment had a practiced fluency and elegance in the writing of the calligraphic shapes. For better or worse, we have been invited to think of this work as a text with a meaning.

For the interested reader who feels motivated or even impelled to continue the search for meaning in the Voynich manuscript, here is a strategy.

The reader would be well advised to return to the original document. Images are available in high resolution, folio by folio, at the website of the Beinecke Rare Book and Manuscript Library. Every line, every string and every glyph can be isolated, studied and compared with known scripts and known languages.

The first decision to be made is: what is the objective?

This author would answer as follows. If the Voynich manuscript or any part of it has meaning, then that meaning must be expressible in a natural human language. It does not matter whether the language is medieval or modern, or what language it is. If the manuscript says something in Latin, it will say something in any language.

The text which expresses meaning must be long enough to be more than a product of chance or selection. A few researchers have claimed to have translated a word of the manuscript, or several words, or even several lines. Even the v101 transliteration, which does not claim to express any meaning, contains by chance some English words such as "okay." But as we have illustrated in previous chapters, with sufficient degrees of freedom we can make sense from nonsense. If there is meaning in the Voynich manuscript, there must be meaningful continuous passages in a single language; and in this author's view, such a passage must represent a least a full page of the manuscript.

The second decision to be made is: what is a glyph?

All of the researchers who have made machine-readable transliterations of the manuscript have had to make this decision. None of them necessarily takes us in the right direction. It is reasonable to suspect that some shapes that have been proposed as single glyphs are in fact strings of two, three or four glyphs; that the "gallows glyphs" are ornamentations of simpler glyphs, in the style of some medieval writers; that each of the "pedestal glyphs" is one glyph sitting on another, or that the pedestal is just ornamentation; also, that some shapes proposed as distinct glyphs are in fact variations of the same glyph. The reader may wish to build a transliteration alphabet, which he or she can base on v101 (as this author has done), or alternatively to construct an entirely new one.

Having done so, it is still possible to use the existing machine-readable transliterations, applying the "find and replace" functions in Microsoft Excel or Word to create a new transliteration (if the reader is so inclined). This is one of the avenues that this author pursued with the alphabets that we have called v102, v103, and v104. For those readers who are content to use the v101 alphabet, or our v104, our appendixes provide tools for experimenting with transliterations. It should not be supposed that glyph and letter frequencies provide a precise key to the transliteration; there is plenty of room for swapping glyphs and letters with similar frequencies.

Once the reader has a machine-readable transliteration, there is a great temptation to do statistical analysis of the text. Researchers have been doing this for decades, and it has not led anywhere. With great caution, and with a view to a mapping to natural languages, this author has done statistical analyses of glyph frequencies and letter frequencies, and their correlations. At best, this approach probably allows us to prioritize the target languages, and to prioritize the glyph-to-letter mappings. For example, this kind of analysis tells us that the target language is more likely to be Italian or Latin than Belarussian or Hebrew.

In this author's view, the prospective researcher would be well advised to suspect or even assume that some glyphs or glyph strings have no meaning in a linguistic sense: that is, that they are punctuation, or markers of some kind. In Chapter 7 we presented evidence, which we feel is compelling, that at least the initial [Φ], the initial [Δ] and the penultimate [8], and probably several other initial and final glyphs, are markers and do not map to any letter in a natural language. If this is so, the text must be cleaned up, or at least the markers have to be clearly identified, prior to any attempt at transliteration to a natural language.

A third strategic decision is to develop an assumption or hypothesis about what constitutes the building blocks of the Voynich manuscript. Practically all modern researchers on the manuscript, from 1912 to the present day (with the courageous exception of Mary D'Imperio), have assumed that a single glyph (however defined) was a building block. This assumption leads swiftly to the expectation that a glyph can map to a letter in a modern alphabet, whether Latin, Greek, Cyrillic, or another.

But this assumption is not necessarily valid. Prescott Currier observed the pervasive links between glyphs and their positions in the glyph strings: for example, the phenomenon that certain glyphs are nearly always initial glyphs, and others are nearly always terminal glyphs. He came close to saying that, if the glyphs had any meaning, it depended on their position in the string. To put this another way: a given glyph could map to more than one letter in a natural language. Likewise, Mary D'Imperio in her technical articles advanced her belief that there was no unique mapping between glyphs and letters in a modern natural language.

This author briefly explored the possibility that glyph pairs, or triplets, or more, could be the building blocks of the Voynich manuscript. This is a

rich field for research; but it will involve decisions about how to parse the glyph strings. It is as if we saw a wall built with bricks of different lengths; but the individual bricks were painted in multicolored sections of equal length, and we could not see where one brick ended and another began.

A fourth strategic decision, if natural languages are the focus of the research, is to prioritize the languages to be studied. The distance-decay theory encourages us to believe that Latin and Italian should be among the top priorities; that other Romance languages, and maybe Bohemian and Old Albanian, should be in the next concentric circle; and that languages that are geographically more distant from Italy should be in the third circle. Our statistical analyses have supported this belief.

If we are looking for languages, and we make the reasonable assumption that the glyphs were written in the fifteenth century, we should be looking at the languages as written in medieval times, rather than the languages as written today. This author has worked on Old Albanian, Middle English, Middle French, Galician-Portuguese, medieval Italian, and medieval Latin, but corpora of other medieval languages are hard to find. That could also be a productive area of research.

Fifthly, if the reader wishes to search for a crib, there is a decision to be made, about where to look. This author has proposed an examination of the "herbal" pages, with the idea that some plant or other is recognizable and that one of its vernacular names, in some language, appears somewhere on the page; but it must be said that not one plant has been identified with confidence. The "zodiac" pages, on the other hand, indisputably identify the constellations to which they refer; possibly the names of some identifiable stars are concealed in these pages. In all the other thematic sections of the manuscript, the quest for a crib is a road that many researchers have traveled without success.

Finally, the reader may well wish to consider whether the manuscript is prose, or poetry, or has some kind of poetic structure wherein repetition, alliteration and rhyming are pervasive. This author has a strong suspicion that the manuscript is a kind of epic poem, and that possibly it draws in some way on the tradition of the Galician-Portuguese *Cantigas*; but only a scholar of the medieval Galician-Portuguese language could take this hypothesis further.

With that, our story comes to an end. We do not claim to have broken the bars of the Voynich manuscript, but perhaps we have rattled the cage. We invite our readers to take up the tools that we have fashioned, or to build new tools of their own; and to continue the journey along this extraordinary road.

EARLY BIRD BOOKS
FRESH DEALS, DELIVERED DAILY

Love this book?

Choose another ebook on us from a selection of similar titles!

- or -

Not loving this book?

No worries - choose another ebook on us from a selection of alternate titles!

CLAIM YOUR FREE EBOOK

AFTERWORD

THE USUAL SUSPECTS

Agatha Christie, in her stories of the adventures of Hercule Poirot, frequently employed a literary device in which the famous Belgian detective, at the culmination of his investigation, would assemble the assorted suspects, preferably at short notice and at uncomfortably close quarters. Thereupon he would lecture each of them in turn, with aplomb and precision, with regard to his certain knowledge that the suspect had had the motive, the means and the opportunity to commit the crime. The suspects would face his discourse with defiance, with insouciance or with guilt, according to their predilections. Finally, he would unmask the true and only perpetrator,[1] and the case would be closed.

We would quite like to close our story of the Voynich manuscript with a Herculean unmasking of the perpetrator; but our case is not so easy.

In the first case, we cannot be sure that there has been a crime. True, there are propositions that the Voynich manuscript was a hoax or forgery, committed for financial gain or for the favor of a patron. But such a proposition involves an assumption about the motives of the author. We have, at best, a single thirdhand report that the manuscript was once sold to

a wealthy collector (namely, the emperor Rudolf II) for a large sum of money; we do not know whether this story is true, and even if it were true, we would not know whether there was anything shady about the transaction. The fact that no one has deciphered the document does not make it a hoax. Likewise, the fact that one can create statistically similar documents by means of an algorithm does not prove that this document was created by throwing dice.

Suppose then that the manuscript is an honest document which possesses meaning, and that the task of Inspector Poirot is simply to determine whodunnit.

In this case, we have a real paucity of perpetrators. The greatest suspect of all, the English philosopher Roger Bacon, was fingered by Wilfrid Voynich himself, and the finger of suspicion remained at least until 1969 as a tentative attribution to Bacon in the catalog of the bookseller Hans Peter Kraus. But we know now that Roger Bacon could not have been the author. He passed away over a hundred years before the creation of the Voynich parchment. We could of course propose that he wrote an earlier version which has been lost, and which was copied by someone else, over a hundred years after his death, onto parchment dating from the fifteenth century. But we could make the same claim about any author who lived prior to the fifteenth century; and we would be straying from the principle of Occam's razor.

Other than Bacon, it is hard to make a case for any medieval author who is a familiar name to modern scholars.

Our fictional detective might reasonably rephrase the question: What kind of person might have produced or commissioned the Voynich manuscript?

We recall from our prologue that the manuscript was a bespoke project: it was one of a kind. Someone had enough money, or could borrow enough money, to assemble and pay for the labor, materials, and organization to bring the manuscript into being.

Therefore, we can add precision to Hercule Poirot's question: What did the manuscript cost to produce, and who had access to that kind of money?

By way of a European example: in 1374 Étienne de Conty, the abbot of Corbie Abbey in Picardy, France, commissioned the scribe Guillaume du Breuil to produce a copy of *In quinque decretalium libros commentaria*, a

book on canon law by Henri Bohic. The book consisted of two volumes, totalling 758 pages; for this work, de Conty paid 62.55 Parisian livres.

The archives of Corbie Abbey record that of this sum, the scribe's salary accounted for 31.25 livres, the parchment (including repairs) for 20.90 livres, illuminated letters for 4.80 livres, and the binding for 1.60 livres.[2]

Carla Bozzolo and Ezio Ornato provide more background.[3] Based on their study of 3,484 manuscripts in medieval French libraries, they estimated that the average value of French manuscripts was 5.35 Parisian livres in the fourteenth century and 2.25 Parisian livres in the fifteenth century. By those standards, the Corbie Abbey manuscript was a massive undertaking.

Finally, by kind courtesy of Dr. Aidan Conti of the University of Bergen, we have assembled a collection of seventy-two medieval European colophons (manuscript annotations) in which the scribes set down their names and the nature of their work. In thirty-nine cases, they recorded what the manuscript had cost or what they personally had received. Here are some examples:

- … pro quarum scriptura valente ducato soldis 100, recepi libras 14 soldos 12.[4]

[… for such a writing worth a sum of 100 ducats (?), I received 10 livres and 12 sous.] (Cremona, 1421)

- Expense novi missalis … Et in feodo Thome Preston per duos annos scribentis dictun missale 4 li. In panno empto pro liberata dicti Thome per dictum tempus 20 s. Summe 33 li. 14 s. 7. d.[5]

[The expenses of the new missal … and in the fee of Thomas Preston for two years writing the said missal 4 pounds. In cloth (clothing?) bought for delivery to the said Thomas in the said time 20 shillings. Total 33 pounds 14 shillings 7 pence.] (London, 1383/1384)

- Lan 1425 … fuit escrip cest presens livre et ung autre parelle, les quels fit faire dame Alix de Villesorreston doienne de remeremont, et li costirent 95 florins.[6]

["In the year 1425 were written the present book and another the same, which were made for Alice de Villesorreston the dame of

Rememont, and they cost 95 (Florentine?) florins."]

For examples from England, Joanne Overty gives a specific case of two illuminated antiphoners, of the highest quality, created in the late fourteenth century for the Chapel of St. Thomas Becket on London Bridge.[7] The materials for the binding alone included two deerskins and ten goatskins plus eight large knots with tassels, at a cost of eighteen shillings. The cost of labor was over thirteen pounds.

We cannot realistically convert these sums from one currency to another, nor from the purchasing power of the medieval era to that of today, but surely we may conclude that writing a medieval manuscript was a work of months or years and was paid accordingly.

The Voynich manuscript is about one-third as long as the Corbie Abbey volumes and has no illuminations. Had it been produced in that time and place (and simply copied from an existing document), it might have cost about ten livres for labour (assuming that the rough-and-ready illustrations were no more expensive to produce than the text), about seven livres for parchment, and one livre for binding: say eighteen livres in all. This cost does not take account of the creative work of authorship, which we have no way to quantify.

To have an approximate idea of how much eighteen livres would buy: Frances Gies and Joseph Gies give an example from Troyes, Champagne in the year 1250, where a quarry mason would have to work for eight to ten weeks to earn a livre in cash, plus board and lodging.[8] To what extent wages in France would have changed in the following century and a quarter, we have no easy way to know.

The Voynich manuscript is believed to have used around fifteen calfskins; we hazard a guess that calfskin was more expensive than deerskin or goatskin. So had the manuscript been produced in England, the skins alone could have cost at least one pound. It was not a work of great quality: perhaps we could allow five pounds for labor and a total budget of seven to ten pounds.

From the classic study by E. H. Phelps Brown and Sheila V. Hopkins,[9] covering seven centuries of wages on English building sites, we know that in mid-fifteenth-century England, an unskilled construction laborer would earn an average of four pence per day, and a construction craftsman an

average of six pence per day. The Voynich manuscript, if produced in England, might therefore have cost anywhere up to the equivalent of 600 days of a laborer's wages.

To express this sum in terms of today's money: in California in 2021, the minimum wage for jobs with small employers was thirteen dollars per hour. If the Voynich manuscript were produced in California today and the relative prices of materials and labor had not changed in the intervening six hundred years, the commissioning party might expect to pay up to around US$60,000.

Whether from the French or from the English examples, we can infer that the producer of the Voynich manuscript must have been quite wealthy. If we use the example from Corbie Abbey, he or she could afford to pay for a single book what a skilled craftsman might take three or four years to earn. Such a person could clearly be a high clerical official, like Étienne de Conty; but we would have to ponder why such a person would commission an undecipherable document. Or the producer could have been a landed aristocrat, or an official at a princely or royal court (these functions would often overlap). In mid-fifteenth-century Europe, there were at least forty kingdoms, duchies, and principalities large enough to be considered sovereign states; we suppose that each of these could have had several hundred to several thousand courtiers at the ruler's palace.

Or the producer could have been a person of independent means, with an enquiring mind, a boundless imagination, and perhaps a colossal sense of humor.

More than this, we cannot say. Much as we would like to hear Inspector Poirot's conclusion, there are too many people who could have done the job. Perhaps one of our readers will identify the language, place, date, and even the author, of the Voynich manuscript, and thereby will be the detective who will bring the story to its conclusion.

APPENDIXES

Author's note: These appendixes are intended to provide tools for the interested reader to map any word, line, paragraph, or page of the Voynich manuscript to a text string in any of the listed medieval languages. The mappings may not, at the first attempt, produce recognisable words. The reader is encouraged to try variations of the correspondences between Voynich glyphs and European letters.

For expanded versions of these appendixes, and for correspondences to languages not listed here, the reader may contact the author through Schiffer Publishing.

Appendix 1A. Voynich manuscript, pages in Language A, v101 ④ transliteration: frequenci common glyphs; and letter frequencies in selected medieval European languages

Voynich manuscript, Language A				medieval languages: sources and correlations							
				Middle English, c.1340		Galician-Portuguese, 1220-1300		Middle French, c.1456		Old Albanian, 1555	
glyph	v101④ key	vowel probability score	% of glyphs	*Auchinleck MS, l.1-2330*		*Cantigas d'Amigo*		*La Farce de Maistre Pathelin*		*Meshari*	
			RSQ=>	95.9%		95.8%		94.9%		94.4%	
o	o	10	17.6%	E	14.6%	E	14.7%	E	17.4%	E	22.0%
9	9	8	10.3%	O	7.9%	A	10.4%	A	7.8%	I	9.1%
a	a	9	8.2%	I	7.5%	O	9.3%	I	7.0%	A	6.1%
c	c	6	6.1%	A	6.8%	I	6.3%	O	6.1%	U	4.4%
to	④	7	2.5%	U	3.3%	U	3.5%	U	5.2%	O	4.1%
ct	1	-	9.1%	N	6.3%	R	7.1%	S	8.3%	T	9.8%
g	e	-	7.0%	S	5.8%	S	6.8%	N	7.1%	N	7.1%

ﻉ	8	-	6.5%	T	5.8%	N	6.6%	R	6.5%	H	6.4%
ﻩ	h	-	5.4%	R	5.6%	M	5.1%	T	6.2%	ſ	4.1%
ﻯ	y	-	5.0%	Þ	4.6%	D	4.6%	L	4.9%	M	3.9%
ﻫ	k	-	3.8%	H	4.5%	L	2.5%	M	3.2%	R	3.8%
ﺎﺍ	m	-	3.3%	D	4.3%	Q	2.4%	D	2.9%	L	2.2%
ﺲ	s	-	2.4%	L	4.1%	V	2.4%	P	2.8%	P	2.2%
ﺣ	2	-	2.4%	W	3.0%	T	2.3%	C	2.4%	C	2.1%
ﺨﻫ	K	-	1.3%	M	2.7%	P	2.1%	V	2.4%	D	2.0%
ﺼ	C	5	1.0%	F	2.3%	C	1.8%	Y	1.6%	ħ	1.7%
ﻐ	p	-	0.8%	C	2.0%	G	1.7%	Q	1.5%	ε	1.3%
ﺨﻩ	H	-	0.7%	B	1.6%	H	1.3%	B	1.4%	B	1.3%
ﻉ	7	-	0.7%	G	1.4%	F	1.2%	J	1.1%	G	1.1%
ﺎﻭ	n	-	0.6%	Y	1.3%	B	0.8%	G	1.1%	F	0.8%

Source: author's analysis. v101④ refers to v101 transliteration with 4o replaced by ④. Pag
defined by Dr Lisa Fagin Davis (2021). Vowel probability scores are based on the Sukhot
by Dr Mans Hulden. RSQ refers to correlation between glyph frequencies and European let

Appendix 1B. Voynich manuscript, pages in Language B, v101 ④ transliteration: frequenci
common glyphs; and letter frequencies in selected medieval European languages

Voynich manuscript, Language B				medieval languages: sources and correlations							
				Italian, medieval		Latin, medieval		Galician-Portuguese, 1220-1300		Middle French, c.1456	
glyph	v101④ key	vowel probability score	% of glyphs	*OVI corpus*		*LatinISE corpus*		*Cantigas d'Amigo*		*Auchinleck MS, l.1-2330*	
			RSQ=>	96.5%		92.8%		89.8%		82.8%	
ﻭ	9	7	11.9%	E	11.6%	E	12.6%	E	14.7%	E	14.6%
ﻙ	c	10	10.1%	A	10.5%	I	11.1%	A	10.4%	O	7.9%
ﻭ	o	8	10.0%	I	9.8%	A	7.4%	O	9.3%	I	7.5%
ﻯ	a	9	9.7%	O	9.4%	U	6.9%	I	6.3%	A	6.8%

ſo	④	6	4.3%	U	2.5%	O	5.9%	U	3.5%	U	3.3%
ﻩ	8	-	7.1%	N	6.6%	T	8.6%	R	7.1%	N	6.3%
ﻫ	h	-	7.1%	R	6.5%	S	8.3%	S	6.8%	S	5.8%
ﻭ	e	-	7.0%	T	5.9%	N	6.4%	N	6.6%	T	5.8%
ﻕ	1	-	6.5%	L	5.8%	M	6.0%	M	5.1%	R	5.6%
ﻯ	y	-	4.1%	S	5.6%	R	4.2%	D	4.6%	Þ	4.6%
ﻑ	k	-	3.8%	C	5.2%	D	3.9%	L	2.5%	H	4.5%
ﻥ	m	-	2.5%	D	3.7%	C	3.5%	Q	2.4%	D	4.3%
ﻕ	C	5	2.2%	M	3.0%	Q	2.8%	V	2.4%	L	4.1%
ﻩ	7	-	2.2%	P	2.8%	L	2.6%	T	2.3%	W	3.0%
ﻕ²	2	-	2.2%	V	2.0%	P	2.5%	P	2.1%	M	2.7%
ﻥ	n	-	1.5%	G	1.9%	B	1.2%	C	1.8%	F	2.3%
ﺭ	s	-	1.3%	H	1.8%	H	1.1%	G	1.7%	C	2.0%
ﻑ	g	-	0.7%	F	1.3%	F	0.6%	H	1.3%	B	1.6%
ﻍ	p	-	0.6%	Q	0.8%	G	0.6%	F	1.2%	G	1.4%
ﻕﻫ	H	-	0.5%	B	0.8%	V	0.6%	B	0.8%	Y	1.3%

Source: author's analysis. v101④ refers to v101 transliteration with 4o replaced by ④. Page defined by Dr Lisa Fagin Davis (2021). Vowel probability scores are based on the Sukhotin by Dr Mans Hulden. RSQ refers to the correlation between glyph frequencies and Europea

Appendix 1C. Voynich manuscript, pages in Language A, v104 transliteration: frequencies common glyphs; and letter frequencies in selected medieval European languages

Voynich manuscript, Language A				medieval languages: sources and correlations v				
				Galician-Portuguese, 1220-1300	Latin, medieval	Middle English, c.1340	Middle French, c.1456	
glyph	v104 key	vowel probability score	% of glyphs	*Cantigas d'Amigo*	*LatinISE corpus*	*Auchinleck MS, l.1-2330*	*La Farce de Maistre Pathelin*	
			RSQ=>	96.8%	93.4%	91.8%	91.6%	
o	o	10	16.3%	E 14.7%	E 12.6%	E 14.6%	E 17.4%	E

ꝛꝛ	1	6	11.1%	A	10.4%	I	11.1%	O	7.9%	A	7.8%	A
୨	Φ	8	9.4%	O	9.3%	A	7.4%	I	7.5%	I	7.0%	I
ɑ	a	9	7.4%	I	6.3%	U	6.9%	A	6.8%	O	6.1%	O
ℐ	Ω	7	3.8%	U	3.5%	O	5.9%	U	3.3%	U	5.2%	U
ᚠ	h	-	10.2%	R	7.1%	T	8.6%	N	6.3%	S	8.3%	N
ȣ	8	-	6.6%	S	6.8%	S	8.3%	S	5.8%	N	7.1%	R
ɾ	e	-	6.3%	N	6.6%	N	6.4%	T	5.8%	R	6.5%	T
ʔ	Ψ	-	4.9%	M	5.1%	M	6.0%	R	5.6%	T	6.2%	L
ᴄ	c	-	4.5%	D	4.6%	R	4.2%	Þ	4.6%	L	4.9%	S
ᵢᵢ	î	-	3.2%	L	2.5%	D	3.9%	H	4.5%	M	3.2%	O
ϯo	Δ	5	2.3%	Q	2.4%	C	3.5%	D	4.3%	D	2.9%	D
ƨ	s	-	2.2%	V	2.4%	Q	2.8%	L	4.1%	P	2.8%	M
ᵢ	i	-	1.9%	T	2.3%	L	2.6%	W	3.0%	C	2.4%	P
ꝛꝛ	ĉ	2	1.7%	P	2.1%	P	2.5%	M	2.7%	V	2.4%	V
ᚠ	g	-	0.9%	C	1.8%	B	1.2%	F	2.3%	Y	1.6%	O
ȣ	Π	4	0.9%	G	1.7%	H	1.1%	C	2.0%	Q	1.5%	H
ᚠ	f	-	0.3%	H	1.3%	F	0.6%	B	1.6%	B	1.4%	P
ϯ	4	3	0.1%	F	1.2%	G	0.6%	G	1.4%	J	1.1%	O
-	£	1	0.0%	B	0.8%	V	0.6%	Y	1.3%	G	1.1%	E

Source: author's analysis. Pages in Language A are as defined by Dr Lisa Fagin Davis (20 scores are based on the Sukhotin algorith; Python code by Dr Mans Hulden. RSQ refers t glyph frequencies and European letter frequencies.

Appendix 1D. Voynich manuscript, pages in Language B, v104 transliteration: frequencies common glyphs; and letter frequencies in selected medieval European languages

Voynich manuscript, Language B				medieval languages: sources and correlations w				
				Latin, medieval	Italian, medieval	Galician-Portuguese, 1220-1300	Middle English, c.1340	
glyph	v104 key	vowel probability	% of glyphs	*OVI corpus*	*LatinISE corpus*	*Cantigas d'Amigo*	*Auchinleck MS, l.1-*	

		score										
			RSQ=>	93.1%		91.8%		85.9%		78.1%		
ϙ	Φ	7	11.0%	E	11.6%	E	12.6%	E	14.7%	E	14.6%	E
ο	o	8	10.6%	A	10.5%	I	11.1%	A	10.4%	O	7.9%	A
ϲ	c	9	9.9%	I	9.8%	A	7.4%	O	9.3%	I	7.5%	I
ɑ	a	10	9.3%	O	9.4%	U	6.9%	I	6.3%	A	6.8%	O
ϑ	Ω	6	3.6%	U	2.5%	O	5.9%	U	3.5%	U	3.3%	U
ϲτ	1	4	9.5%	N	6.6%	T	8.6%	R	7.1%	N	6.3%	S
ϐ	8		8.2%	R	6.5%	S	8.3%	S	6.8%	S	5.8%	N
ᾑ	h		6.9%	T	5.9%	N	6.4%	N	6.6%	T	5.8%	R
ϱ	e		6.5%	L	5.8%	M	6.0%	M	5.1%	R	5.6%	T
ϼ	Ψ		4.4%	S	5.6%	R	4.2%	D	4.6%	Þ	4.6%	L
ᾗ	k		4.0%	C	5.2%	D	3.9%	L	2.5%	H	4.5%	M
ϯο	Δ	5	3.5%	D	3.7%	C	3.5%	Q	2.4%	D	4.3%	D
ιι	î		2.5%	M	3.0%	Q	2.8%	V	2.4%	L	4.1%	R
ϲϲ	ĉ	3	2.0%	P	2.8%	L	2.6%	T	2.3%	W	3.0%	O
ι	i		1.9%	V	2.0%	P	2.5%	P	2.1%	M	2.7%	V
ϩ	s		1.5%	G	1.9%	B	1.2%	C	1.8%	F	2.3%	Y
ᾔ	g		1.0%	H	1.8%	H	1.1%	G	1.7%	C	2.0%	O
ϐ	Π	2	0.7%	F	1.3%	F	0.6%	H	1.3%	B	1.6%	B
ᾕ	f		0.3%	Q	0.8%	G	0.6%	F	1.2%	G	1.4%	J
ϯ	4		0.1%	B	0.8%	V	0.6%	B	0.8%	Y	1.3%	O

Source: author's analysis. Pages in Language A are as defined by Dr Lisa Fagin Davis (20 scores are based on the Sukhotin algorith; Python code by Dr Mans Hulden. RSQ refers t glyph frequencies and European letter frequencies.

Appendix 2A. Voynich manuscript, pages in Language A, v101 ④ transliteration: frequenci twenty most common glyph pairs; and bigram frequencies in selected medieval European la

Voynich manuscript, Language A			medieval languages, sources and correlations with Lang									
			Latin, medieval		Early New High German, c.1401		Middle English, c.1340		Albanian, 1555		Italian, medieval	
glyphs	v101④ key	% of 2-grams	*LatinISE medieval subcorpus*		*Der Ackermann aus Böhmen*		*Auchinleck MS, lines 1-2330*		*Meshari, Ch 1-140*		*OVI corpus*	
		RSQ=>	97.9%		95.9%		95.3%		93.5%		91.8%	
o9	oe	3.9%	QU	2.7%	EN	3.3%	HE	1.6%	TE	3.0%	ER	2.1%
cco	1o	3.2%	ET	2.1%	ER	3.1%	ÞE	1.4%	EH	2.7%	ON	1.7%
88	8a	2.4%	UM	1.8%	CH	2.3%	ER	1.4%	NE	2.4%	RE	1.6%
a11J	am	2.3%	IN	1.7%	DE	1.9%	OU	1.3%	EN	2.3%	DI	1.6%
oy	oy	2.3%	ES	1.6%	EI	1.7%	DE	1.3%	ER	1.5%	NO	1.6%
co	co	2.0%	ER	1.5%	GE	1.5%	EN	1.1%	SE	1.4%	AN	1.6%
ccc	1c	1.6%	US	1.4%	IN	1.4%	ÞA	1.0%	ME	1.3%	DE	1.5%
olf	oh	1.5%	ST	1.3%	TE	1.4%	HI	0.9%	ET	1.3%	CO	1.4%
o8	o8	1.5%	UI	1.2%	ND	1.2%	CH	0.9%	TI	1.1%	EN	1.4%
c9	c9	1.3%	TE	1.2%	HE	1.2%	RE	0.9%	IT	1.0%	CH	1.4%
89	89	1.3%	IT	1.2%	LE	1.1%	IN	0.9%	IE	0.9%	TO	1.4%
a9	ae	1.3%	AM	1.1%	ES	1.1%	LE	0.9%	TU	0.8%	RA	1.3%
olf	ok	1.2%	SI	1.1%	IC	1.0%	TO	0.9%	AT	0.8%	TE	1.3%
ay	ay	1.2%	DE	1.1%	RE	1.0%	AT	0.8%	CH	0.8%	LA	1.3%
cc9	19	1.0%	SE	1.1%	BE	1.0%	IS	0.8%	⊦I	0.8%	EL	1.2%
c̊co	2o	0.9%	IS	1.0%	VN	1.0%	ON	0.8%	ND	0.7%	LI	1.2%
⊦olf	④h	0.9%	NO	1.0%	ST	0.9%	TE	0.8%	PE	0.7%	OR	1.2%
lfc	hc	0.9%	UA	1.0%	AN	0.9%	AN	0.7%	AN	0.7%	AL	1.2%
ffcc	k1	0.9%	EM	0.9%	WE	0.8%	OR	0.7%	IH	0.7%	IN	1.2%
lfcc	h1	0.8%	ON	0.9%	IE	0.8%	WE	0.7%	ƐO	0.7%	AR	1.1%

Source: author's analysis. v101 ④ refers to v101 transliteration with 4o replaced by Language A are as defined by Dr Lisa Fagin Davis (2021). Bigram frequencies (exclu

calculated using dcode.fr/frequency-analysis. RSQ refers to correlation between frequer pairs and frequencies of European bigrams.

Appendix 2B. Voynich manuscript, pages in Language B, v101 ④ transliteration: frequenci twenty most common glyph pairs; and bigram frequencies in selected medieval European la

Voynich manuscript, Language B			Latin, medieval		Italian, medieval		Middle English, c.1340		Galician-Portuguese, c.1220-1300		Middle French, c.1456	
glyphs	v101④ key	% of 2-grams	*LatinISE medieval subcorpus*		*OVI corpus*		*Auchinleck MS, lines 1-2330*		*Cantigas d'Amigo*		*La Farce de Maistr Pathelin*	
		RSQ=>	97.2%		97.0%		96.2%		94.4%		93.6%	
ꝺ9	89	3.3%	QU	2.7%	ER	2.1%	HE	1.6%	QU	1.7%	EN	1.8%
ꝯc	1c	2.7%	ET	2.1%	ON	1.7%	ÞE	1.4%	DE	1.6%	ES	1.7%
cꝺ	c8	2.4%	UM	1.8%	RE	1.6%	ER	1.4%	ER	1.5%	OU	1.6%
oꝛ	oe	2.4%	IN	1.7%	DI	1.6%	OU	1.3%	EN	1.5%	RE	1.5%
ꝑoꝰ	④h	2.2%	ES	1.6%	NO	1.6%	DE	1.3%	UE	1.4%	QU	1.2%
ꝰa	ha	2.1%	ER	1.5%	AN	1.6%	EN	1.1%	OR	1.3%	IE	1.1%
aꝛ9	am	1.8%	US	1.4%	DE	1.5%	ÞA	1.0%	EU	1.2%	NT	1.1%
aꝶ	ay	1.8%	ST	1.3%	CO	1.4%	HI	0.9%	SE	1.2%	AI	1.1%
aꝛ	ae	1.7%	UI	1.2%	EN	1.4%	CH	0.9%	MI	1.1%	LE	1.0%
ꝰc	hc	1.5%	TE	1.2%	CH	1.4%	RE	0.9%	ON	1.1%	ST	1.0%
c9	c9	1.4%	IT	1.2%	TO	1.4%	IN	0.9%	ES	1.1%	DE	1.0%
ꝺ9	79	1.4%	AM	1.1%	RA	1.3%	LE	0.9%	AN	1.0%	VO	0.9%
ꝺa	8a	1.4%	SI	1.1%	TE	1.3%	TO	0.9%	RE	0.9%	US	0.9%
oꝰ	oh	1.3%	DE	1.1%	LA	1.3%	AT	0.8%	AR	0.9%	ON	0.9%
oꝓꝰ	ok	1.2%	SE	1.1%	EL	1.2%	IS	0.8%	RA	0.9%	ER	0.9%
aꝛꝺ	an	1.1%	IS	1.0%	LI	1.2%	ON	0.8%	ME	0.8%	NE	0.9%
ꝯc	2c	1.1%	NO	1.0%	OR	1.2%	TE	0.8%	PO	0.8%	ME	0.8%
cꝺ	c7	1.1%	UA	1.0%	AL	1.2%	AN	0.7%	OS	0.8%	TE	0.8%
co	co	1.0%	EM	0.9%	IN	1.2%	OR	0.7%	EI	0.8%	SE	0.7%

ꞓꞓ	cc	1.0%	ON	0.9%	AR	1.1%	WE	0.7%	AM	0.7%	IL	0.7

Source: author's analysis. v101 ④ refers to v101 transliteration with 4o replaced by Language B are as defined by Dr Lisa Fagin Davis (2021). Bigram frequencies (excl calculated using dcode.fr/frequency-analysis. RSQ refers to correlation between freque pairs and frequencies of European bigrams.

Appendix 2C. Voynich manuscript, pages in Language A, v104 transliteration: frequencies twenty most common glyph pairs; and bigram frequencies in selected medieval European la

Voynich manuscript, Language A			medieval languages, sources and correlations with Langua									
			Galician-Portuguese, 1220-1300		Latin, medieval		Early New High German, c.1401		Italian, medieval		Middle French, c.1456	
glyphs	v104 key	% of 2-grams	*Cantigas d'Amigo*		*LatinISE medieval subcorpus*		*Der Ackerman n aus Böhmen*		*OVI corpus*		*La Farce de Maistre Pathelin*	
		RSQ=>	92.7%		92.0%		91.6%		91.1%		89.7%	
ꞓꞓo	1o	3.7%	QU	1.7%	QU	2.7%	EN	3.3%	ER	2.1%	EN	1.8%
oꝶ	oe	3.7%	DE	1.6%	ET	2.1%	ER	3.1%	ON	1.7%	ES	1.7%
ꝺa	8a	2.6%	ER	1.5%	UM	1.8%	CH	2.3%	RE	1.6%	OU	1.6%
ıı℩	îΩ	2.5%	EN	1.5%	IN	1.7%	DE	1.9%	DI	1.6%	RE	1.5%
ꞓ̓	1’	2.4%	UE	1.4%	ES	1.6%	EI	1.7%	NO	1.6%	QU	1.2%
ɑ ıı	aî	2.3%	OR	1.3%	ER	1.5%	GE	1.5%	AN	1.6%	IE	1.1%
oʔ	oψ	2.3%	EU	1.2%	US	1.4%	IN	1.4%	DE	1.5%	NT	1.1%
ꞓo	co	2.0%	SE	1.2%	ST	1.3%	TE	1.4%	CO	1.4%	AI	1.1%
ꟻꞓꞓ	k1	2.0%	MI	1.1%	UI	1.2%	ND	1.2%	EN	1.4%	LE	1.0%
ꞓꞓꞓ	1c	1.9%	ON	1.1%	TE	1.2%	HE	1.2%	CH	1.4%	ST	1.0%
ꞓꞓ9	1Φ	1.6%	ES	1.1%	IT	1.2%	LE	1.1%	TO	1.4%	DE	1.0%
oꝺ	o8	1.5%	AN	1.0%	AM	1.1%	ES	1.1%	RA	1.3%	VO	0.9%
ꟻꞓꞓ	h1	1.5%	RE	0.9%	SI	1.1%	IC	1.0%	TE	1.3%	US	0.9%
oꟻ	oh	1.5%	AR	0.9%	DE	1.1%	RE	1.0%	LA	1.3%	ON	0.9%

ﻉ	8Φ	1.4%	RA	0.9%	SE	1.1%	BE	1.0%	EL	1.2%	ER	0.9%
ﻉ	cΦ	1.3%	ME	0.8%	IS	1.0%	VN	1.0%	LI	1.2%	NE	0.9%
ﻭﻑ	ok	1.2%	PO	0.8%	NO	1.0%	ST	0.9%	OR	1.2%	ME	0.8%
’ﻭ	’o	1.1%	OS	0.8%	UA	1.0%	AN	0.9%	AL	1.2%	TE	0.8%
ﻉﺭ	aΨ	1.0%	EI	0.8%	EM	0.9%	WE	0.8%	IN	1.2%	SE	0.7%
ﻑﻍ	hc	0.9%	AM	0.7%	ON	0.9%	IE	0.8%	AR	1.1%	IL	0.7%

Source: author’s analysis. Pages in Language A are as defined by Dr Lisa Fagin Davis (2
frequencies (excluding spaces) calculated using dcode.fr/frequency-analysis. RSQ refers t
between frequencies of glyph pairs and frequencies of European bigrams.

Appendix 2D. Voynich manuscript, pages in Language B, v104 transliteration: frequencies
twenty most common glyph pairs; and bigram frequencies in selected medieval European la

Voynich manuscript, Language B			medieval languages, sources and correlations with Langua									
			Latin, medieval		Italian, medieval		Middle English, c.1340		Early New High German, c.1401		Middle French, c.1456	
glyphs	v104 key	% of 2-grams	*LatinISE medieval subcorpus*		*OVI corpus*		*Auchinleck MS, lines 1-2330*		*Der Ackermann aus Böhmen*		*La Farce de Maistre Pathelin*	
		RSQ=>	96.3%		91.5%		89.9%		88.8%		88.3%	
ﻉ	8Φ	4.3%	QU	2.7%	ER	2.1%	HE	1.6%	EN	3.3%	EN	1.8%
ﻉ	c8	2.8%	ET	2.1%	ON	1.7%	ÞE	1.4%	ER	3.1%	ES	1.7%
ﻕﻕ	1c	2.6%	UM	1.8%	RE	1.6%	ER	1.4%	CH	2.3%	OU	1.6%
ﻭﻍ	oe	2.5%	IN	1.7%	DI	1.6%	OU	1.3%	DE	1.9%	RE	1.5%
’ﻕ	1’	2.4%	ES	1.6%	NO	1.6%	DE	1.3%	EI	1.7%	QU	1.2%
ﻉﻍ	ae	1.9%	ER	1.5%	AN	1.6%	EN	1.1%	GE	1.5%	IE	1.1%
ﻉﺭ	aψ	1.8%	US	1.4%	DE	1.5%	ÞA	1.0%	IN	1.4%	NT	1.1%
ﻑﻉ	ha	1.8%	ST	1.3%	CO	1.4%	HI	0.9%	TE	1.4%	AI	1.1%
ﻁﻭﻑ	Δh	1.8%	UI	1.2%	EN	1.4%	CH	0.9%	ND	1.2%	LE	1.0%
ﻉﻉ	8a	1.7%	TE	1.2%	CH	1.4%	RE	0.9%	HE	1.2%	ST	1.0%

ıı9	îΩ	1.7%	IT	1.2%	TO	1.4%	IN	0.9%	LE	1.1%	DE	1.0%
ɑıı	aî	1.6%	AM	1.1%	RA	1.3%	LE	0.9%	ES	1.1%	VO	0.9%
oſ	oh	1.5%	SI	1.1%	TE	1.3%	TO	0.9%	IC	1.0%	US	0.9%
cǫ	cΦ	1.5%	DE	1.1%	LA	1.3%	AT	0.8%	RE	1.0%	ON	0.9%
’c	’c	1.5%	SE	1.1%	EL	1.2%	IS	0.8%	BE	1.0%	ER	0.9%
oſ	ok	1.5%	IS	1.0%	LI	1.2%	ON	0.8%	VN	1.0%	NE	0.9%
ſc	hc	1.4%	NO	1.0%	OR	1.2%	TE	0.8%	ST	0.9%	ME	0.8%
co	co	1.3%	UA	1.0%	AL	1.2%	AN	0.7%	AN	0.9%	TE	0.8%
ɑı	ai	1.2%	EM	0.9%	IN	1.2%	OR	0.7%	WE	0.8%	SE	0.7%
cc	cc	1.1%	ON	0.9%	AR	1.1%	WE	0.7%	IE	0.8%	IL	0.7%

Source: author’s analysis. Pages in Language B are as defined by Dr Lisa Fagin Davis (20
frequencies (excluding spaces) calculated using dcode.fr/frequency-analysis. RSQ refers to
between frequencies of glyph pairs and frequencies of European bigrams.

Appendix 3A. Voynich manuscript, pages in Language A, v104 transliteration: frequencies
common glyph triplets; and trigram frequencies in selected medieval European languages

Voynich manuscript, Language A			medieval languages, sources and correlations with La								
			Galician-Portuguese, 1220-1300		Albanian, 1555		Middle English, c.1500		Early New High German, c.1401		Ita mec
glyphs	v104 key	% of 3-grams	*Cantigas d’Amigo*		*Meshari, Ch 1-140*		*Ashmole 61, parts 1-27*		*Der Ackermann aus Böhmen*		*OVI*
		RSQ=>	97.7%		95.1%		94.8%		93.6%		88
ɑıı9	aîΩ	2.3%	QUE	1.3%	TEH	1.0%	THE	1.6%	ICH	1.0%	CHE
ȣɑıı	8aî	1.3%	MIG	0.8%	ƐOT	0.7%	AND	0.9%	EIN	0.8%	PER
ɑcoɋ	1oe	1.2%	AMI	0.7%	END	0.6%	THA	0.6%	DER	0.7%	ELL
ć2o	1’o	1.2%	IGO	0.6%	PER	0.6%	HAT	0.5%	CHE	0.7%	ENT
ɑcoʔ	1oΨ	0.9%	DES	0.5%	ETE	0.6%	HER	0.5%	VND	0.6%	CON
ſɑco	k1o	0.8%	MEU	0.5%	ENE	0.6%	GHT	0.5%	CHT	0.6%	QUE

⫛⫛o	1co	0.7%	NON	0.5%	MEH	0.5%	THO	0.4%	SCH	0.5%	DEL
o8ϙ	o8Φ	0.7%	VOS	0.4%	UNE	0.5%	YGH	0.4%	DEN	0.5%	QUA
⫛⫛ϙ	1cΦ	0.6%	POR	0.4%	NDE	0.4%	FOR	0.4%	NDE	0.5%	ARE
⫯⫛ϙ	k1Φ	0.6%	BEN	0.3%	SEH	0.4%	HOU	0.4%	LLE	0.5%	AND
⫛̓⫛	1’c	0.6%	AND	0.3%	TUN	0.4%	ALL	0.4%	GEN	0.4%	LLA
⫯⫛o	h1o	0.5%	MOR	0.3%	ERI	0.4%	SCH	0.3%	HEN	0.4%	NTE
o8a	o8a	0.5%	ANT	0.3%	INE	0.4%	WAS	0.3%	ALL	0.3%	EST
⫛o8	1o8	0.5%	REI	0.3%	IET	0.4%	HEI	0.3%	TEN	0.3%	GLI
coϤ	coe	0.5%	ORA	0.3%	UOM	0.3%	HYM	0.3%	WER	0.3%	NON
o⫯⫛	ok1	0.5%	COM	0.2%	ATE	0.3%	ERE	0.2%	IST	0.3%	MEN
⫯⫛ϙ	h1Φ	0.5%	GRA	0.2%	ANE	0.3%	HIS	0.2%	LIC	0.3%	ESS
a\ℐ	aiΩ	0.5%	RAN	0.2%	SSE	0.3%	CHE	0.2%	INE	0.3%	ERE
8a\	8ai	0.4%	CON	0.2%	TUU	0.3%	YLL	0.2%	DIE	0.3%	COM
o⫯⫛	oh1	0.4%	UER	0.2%	NEH	0.3%	THI	0.2%	DAS	0.3%	ERA

Source: author’s analysis. Pages in Language A are as defined by Dr Lisa Fagin D frequencies (excluding spaces) calculated using dcode.fr/frequency-analysis. RSQ refers t frequencies of glyph triplets and frequencies of European trigrams.

Appendix 3B. Voynich manuscript, pages in Language B, v104 transliteration: frequencies common glyph triplets; and trigram frequencies in selected medieval European languages

Voynich manuscript, Language B			medieval languages: sources and correlations with La								
			Latin, medieval		Middle English, c.1340		Galician-Portuguese, c.1220-1300		Early New High German, c.1401		Alba 15
glyphs	v104 key	% of 3-grams	*OVI corpus*		*Auchinleck MS, lines 1-2330*		*Cantigas d’Amigo*		*Der Ackermann aus Böhmen*		*Mesh 1-*
		RSQ=>	94.4%		93.9%		93.6%		87.1%		85
cϞϙ	c8Φ	1.9%	CHE	0.8%	ÞAT	0.6%	QUE	1.3%	ICH	1.0%	TEH
a\\ℐ	aîΩ	1.4%	PER	0.8%	CHE	0.6%	MIG	0.8%	EIN	0.8%	ƐOT

ƈƈc	1’c	1.3%	ELL	0.7%	SCH	0.6%	AMI	0.7%	DER	0.7%	END
ƈƈc8	1c8	0.8%	ENT	0.5%	WAS	0.4%	IGO	0.6%	CHE	0.7%	PER
o8ϑ	o8Φ	0.7%	CON	0.5%	IȝT	0.4%	DES	0.5%	VND	0.6%	ETE
okc	okc	0.7%	QUE	0.5%	FOR	0.4%	MEU	0.5%	CHT	0.6%	ENE
ohc	ohc	0.7%	DEL	0.4%	WIÞ	0.4%	NON	0.5%	SCH	0.5%	MEH
oha	oha	0.6%	QUA	0.4%	ÞER	0.3%	VOS	0.4%	DEN	0.5%	UNE
8aɋ	8ae	0.6%	ARE	0.4%	EUE	0.3%	POR	0.4%	NDE	0.5%	NDE
aıȣ	aiΩ	0.6%	AND	0.4%	HIR	0.3%	BEN	0.3%	LLE	0.5%	SEH
co8	co8	0.6%	LLA	0.4%	LLE	0.3%	AND	0.3%	GEN	0.4%	TUN
haɋ	hae	0.5%	NTE	0.4%	NDE	0.3%	MOR	0.3%	HEN	0.4%	ERI
ƈƈc9	1cΦ	0.5%	EST	0.3%	HIS	0.3%	ANT	0.3%	ALL	0.3%	INE
8aıı	8aÎ	0.5%	GLI	0.3%	OUȝ	0.3%	REI	0.3%	TEN	0.3%	IET
8aʔ	8aΨ	0.5%	NON	0.3%	UȝT	0.3%	ORA	0.3%	WER	0.3%	UOM
ƭoha	Δha	0.5%	MEN	0.3%	ERE	0.3%	COM	0.2%	IST	0.3%	ATE
hc8	hc8	0.4%	ESS	0.3%	ICH	0.3%	GRA	0.2%	LIC	0.3%	ANE
ƈco	1co	0.4%	ERE	0.3%	OUN	0.3%	RAN	0.2%	INE	0.3%	SSE
hco	hco	0.4%	COM	0.3%	HIM	0.3%	CON	0.2%	DIE	0.3%	TUU
oha	oka	0.4%	ERA	0.3%	ING	0.2%	UER	0.2%	DAS	0.3%	NEH

Source: author’s analysis. Pages in Language B are as defined by Dr Lisa Fagin D frequencies (excluding spaces) calculated using dcode.fr/frequency-analysis. RSQ refers t frequencies of glyph triplets and frequencies of European trigrams.

Appendix 4. Selected corpora of medieval European languages

language	corpus	number of words	time period
Czech	*DIAKORP v6*	4,000,000	14th to 20th c.
Czech, German, Latin, Polish	*“PolDiLemma” Middle Polish Diachrone Lemmatised Corpus*	7,000,000	16th to 18th c.
Dutch	*Brieven als buit*	460,000	17th to 18th c.

Dutch	*Corpus Gysseling*	1,500,000	13th c.
English	*A Corpus of English Dialogues 1560-1760 (CED)*	1,200,000	1560 to 1760
English	*Corpus of Early English Correspondence Sampler (CEECS)*	450,000	1418 to 1680
French	*Syntactic Reference Corpus of Medieval French*	245,000	9th to 13th c.
German	*Austrian Baroque Corpus*	200,000	1650 to 1750
German	*GeMi Corpus*	120,000	1500 to 1700
German	*Reference Corpus Middle Low German/Low Rhenish*	200,700	1200 to 1650
German	*Referenzkorpus Mittelhochdeutsch*	2,500,000	1050 to 1350
Greek	*Greek Medieval Texts*	3,400,000	4th c. - 16th c.
Hungarian	*Corpus of Old and Middle Hungarian court records and private correspondence*	850,000	16th to 18th c.
Hungarian	*Old Hungarian Corpus*	3,000,000	12th to 17th c.
Italian	*Corpus OVI dell'Italiano antico*	540,210	before 1400
Italian	*Corpus testuale del Tesoro della Lingua Italiana delle Origini*	23,000,000	before 1375
Italian	*La Divina Commedia, by Dante Alighieri*	100,000	c. 1308 - 1320
Italian	*M.I.DIA. (Morfologia dell'Italiano in DIAcronia)*	7,500,000	13th to 20th c.
Latin	*LatinISE corpus (version 4)*	13,300,000	-2nd to 21st c.
Portuguese	*Corpus Informatizado do Português Medieval*	2,000,000	9th to 16th c.
Portuguese	*Parsed Corpus of Historical Portuguese*	3,300,000	14th to 19th c.
Slovenian	*Digital library and corpus of historical Slovene IMP 1.1*	17,700,000	1584 to 1919

Spanish	*Corpus Diacrónico del Español (CORDE)*	300,000,000	12th c? to 1974

Source: author's research.

APPENDIX 5: THE SUKHOTIN ALGORITHM

In various chapters of this book, the author has proposed that frequency analysis, if applied systematically to a large number of permutations of Voynich transliterations and target languages, might be a basis for identifying meaning within the text. This is a task for programmers with plenty of computing power: more than this author possesses.

However, if we could classify the glyphs into those that represent vowels, and those that represent consonants, our task would be simpler by an order of magnitude. The number of permutations, at least, would be greatly reduced. In that case, we would only need to map vowel-glyphs to vowels, and consonant-glyphs to consonants. Most medieval European languages have about five vowels, and about twenty consonants (in both cases, for the moment, leaving aside accents and diacritics). In the Voynich manuscript, there are at most twenty-five distinct glyphs that account for over 95 percent of the text.

The question is therefore: if we have before us an unknown script, and we make the heroic assumption that the individual glyphs have been mapped from letters in a natural language, is there an algorithm for identifying the vowels and the consonants?

Thanks to the work of the French linguist Jacques Guy, we learn that such an algorithm exists. It was published in 1962 by Boris V. Sukhotin in his article "Экспериментальное видение буквенных классов с помощью EBM" ("Experimental Vision of Letter Classes Using EVM"), in *Проблемы Структурной Лингвистики* (*Problems of Structural Linguistics*).

Guy, in his article "Vowel identification: An old (but good) algorithm" in *Cryptologia* of July 1991, described Sukhotin's algorithm as follows:

> Sukhotin … assume[s] a state of complete ignorance about the language, except that the writing system is alphabetical…. Sukhotin had observed that vowels tend to occur next to consonants rather than next to vowels.

Guy then worked through a manual example of Sukhotin's algorithm, applied to the word SAGITTA, and found that the algorithm correctly identified the vowels as A and I, and the consonants as G, S, and T.

John Goldsmith and Aris Xanthos, in their article "Learning Phonological Categories" (*Language*, March 2009), observed that Sukhotin had made another basic assumption. They wrote:

> Sukhotin's algorithm … relies on two fundamental assumptions: first, that the most frequent symbol in a transcription is always a vowel, and second, that vowels and consonants tend to alternate more often than not."

An extract from a cartoon in Guy's article in *Cryptologia* of July 1991. The caption reads: "The artisan hit his thumb." *Image credit: JD*

Marco Ponzi, a member of the Voynich Ninja forum, has noted that Goldsmith's and Xanthos's interpretation is not quite correct. Sukhotin assumed that the most probable vowel was the symbol that most often occurred adjacent to another symbol. A symbol which is interior to a word

has two adjacent symbols; an initial or final symbol has only one; an isolated symbol has none. Therefore, if the most frequent symbol is initial, final, or isolated, it is not necessarily a vowel.

We may infer that the way the Sukhotin algorithm works is as follows:

- Find the symbol with the most occurrences adjacent to another; this is most probably a vowel.
- The immediately preceding and following symbols (if any) are probably consonants.
- If those symbols occur elsewhere in the text, the immediately preceding and following symbols (if any) are probably vowels.
- And so on, iteratively, until each symbol has been identified as either a probable vowel or a probable consonant.

This is of course a statistical approach, based on probabilities. It does not exclude pairs of vowels. Most phonetic languages have a few relatively common vowel pairs. For example, as we see from Stefan Trost's excellent website, in modern Italian the most frequent vowel pair is "IO", with a frequency of 1.1 percent; in classical Latin, "AE" (1.0 percent); in modern French, "AI" (1.9 percent).

Sukhotin was evidently working in the tradition of the Russian mathematician Andrey Andreyevich Markov, who created the concept now known as Markov chains (as described in Chapter 7).

Sukhotin conceived the algorithm as a manual process, and possibly intended for it to be used on one word at a time. With the advent of cheap and accessible computing, it has become possible to express the algorithm in programming languages such as Python, and to run it on texts of unlimited length.

Dr. Mans Hulden, of the University of Colorado at Boulder, has developed a Python code for implementation of the Sukhotin algorithm. This code was one of the bases for our proposed mappings of Voynich glyphs to European letters, in appendixes 1.1 through 1.4.

Hulden's code has the advantage of listing the characters in descending order of probability that they are vowels. For example, the code, applied to

the pages in Language A in the v101 transliteration, yields the following glyphs as probable vowels:

o {o}, a {a}, 9 {9}, to {4o}, c {c}, a {A}, q {(}, ccc {d}, c {E}, cc {C}

Since we normally expect only five vowels in a European language, we can reasonably assume that the first five glyphs in this series represent vowels, and that the last five glyphs (which are in any case relatively infrequent) either represent consonants or are variations of the first five.

For interested readers, Dr Hulden's Python code can be downloaded from https://www.voynich.ninja/thread-3901-post-56566.html#pid56566. If the code is run in Windows on the command line, it will ask for an input text file, which should be specified as a file path, for example c:\vms\input.txt.

ENDNOTES

Prologue

1. Parchment refers to any animal skin prepared for writing purposes; vellum refers specifically to calf skin and is typically of a higher quality than other types of parchment.

2. Hodgins, Gregory W. L. (2009). "Dating Results for Voynich Manuscript Parchment Leaves 8, 26, 47 and 68, our reference AA84050, 49, 48, 47 respectively" (letter to Christine McCarthy, Chief Conservator for Special Collections, Yale University Library, dated February 17, 2009), published by Richard SantaColoma at https://proto57.wordpress.com/2021/05/14/the-long-awaited-voynich-radiocarbon-report/.

3. Reddy and Knight (2011) wrote: "… the McCrone Research Institute has asserted that the ink was added shortly afterward [after the production of the parchment]." However, there is no published documentation of this assertion.

4. Avakov, Alexander V. (2015). *Two Thousand Years of Economic Statistics, Volume 1*, pp. 12–13.

5. Frucht, Richard C., ed. (2005). *Eastern Europe*. ABC-CLIO. p. 252. ISBN 978-1-57607-800-6.

6. Zandbergen, René (20 March 2021). The Voynich MS – Biographies: B. Jacobus Horčický de Tepenec (Sinapius).

7. Neal, Philip (undated). The letter of Georgius Barschius to Athanasius Kircher (1637) [English translation].

8. German was at the time one of the official languages of Bohemia, but Bohemia was not part of any German state.

9. Neal, Philip (undated). The letter of Athanasius Kircher to Theodorus Moretus (1639) [English translation].

10. Zandbergen, René (2021a). The Voynich MS – Biographies: D. Georgius Barschius (Baresch), last updated March 20, 2021.

11. Neal, Philip (undated). Johannes Marcus Marci, *Philosophia Vetus Restituta* (1662) [English translation].

12. Neal, Philip (undated). The letter of Johannes Marcus Marci to Athanasius Kircher (1665).

13. Zandbergen, René (2021b). The Voynich MS – Biographies: C. Raphael Mnišovský, last updated March 20, 2021.

14. Voynich, Wilfrid M. (1921). A Preliminary Sketch of the History of the Roger Bacon Cipher Manuscript (Philadelphia, *Transactions of the College of Physicians of Philadelphia*, Vol. 43, 1921, pp. 415–430.

15. This sentence is a footnote in the published proceedings of the event.

16. Ibid.

17. Ibid.

18. Zandbergen, René (2020). Manuscripts Voynich acquired in 1911–1912. Last updated September 18, 2020.

19. Zandbergen, René (2018). The History of the Voynich MS, last updated October 10, 2018.

Chapter 1

1. MacKinnon, Colin (2013a). The Naturalization Papers of Wilfrid Michael Voynich, p. 9.

2. Ibid.

3. Irkutsk Regional Museum of Local Lore, "State criminal from the nobility Mikhail Voynich," inventory no. 18-92, cited by Skorokhodov, Vladimir Pavlovich (1957): *Istoricheskiy Arkhiv*, Moscow, July–August 1957, No. 4, "Khronika," VII, pp. 244–245. Author's translation.

4. Skorokhodov (1957), p. 245. Author's translation.

5. Ibid, p. 245. Author's translation.

6. MacKinnon, Colin (2013b). The Voynich "Coat of Arms" and Michael Voynich's Flight from Siberia, p. 4.

7. Skorokhodov (1957), p. 245. Author's translation.

8. Ibid., p. 245. Author's translation.

9. Ibid., p. 245. Author's translation.

10. Ibid., p. 245. Author's translation.

11. MacKinnon, Colin (2013c). The Voyniches in the 1901 Census of England and Wales, p. 2.

12. MacKinnon, Colin (2013d). The registration of marriage of Ethel and Wilfrid Voynich, p. 2.

13. MacKinnon, Colin (2013a). The Naturalization Papers of Wilfrid Michael Voynich, p. 2.

14. Wilfrid Voynich, June 25, 1917, letter to W. S. Booth, reproduced in MacKinnon, Colin, 2013: The US Bureau of Investigation file on Wilfrid Voynich.

15. W. S. Booth (1917). Letter to Department of Justice, July 7, 1917, reproduced in MacKinnon, Colin (2013e) The US Bureau of Investigation file on Wilfrid Voynich.

16. Charles Daniel Frey (1917). Report apparently to BoI file, November 22, 1917, reproduced in MacKinnon, Colin (2013e) The US Bureau of Investigation file on Wilfrid Voynich.

17. A. W. Willett (1917). Report apparently to BoI file, December 18, 1917, reproduced in MacKinnon, Colin (2013e) The US Bureau of Investigation file on Wilfrid Voynich.

18. Prinke, Rafał T. (2012). "History of the Manuscript from Rudolf to Voynich," presentation at the 100th anniversary of the discovery of the Voynich MS, Villa Mondragone, Monteporzio Catone, May 11, 2012.

19. Capt. [John?] Manly (1917). Letter to Lt. Straus, December 27, 1917, reproduced in MacKinnon, Colin (2013e) The US Bureau of Investigation file on Wilfrid Voynich.

20. MacKinnon, Colin (2013f). *The Travels of Ethel and Wilfrid Voynich and Anne M. Nill between the US and the UK According to American and British Records*, pp. 2, 3.

Chapter 2

1. D'Imperio, Mary E. (1978). *The Voynich Manuscript: An Elegant Enigma* (National Security Agency / Central Security Service, Fort George G. Meade, MD), p. 79.

2. Zandbergen, René (2020). History of research of the Voynich.

3. Hodgins, Gregory W. L. (2009). "Dating Results for Voynich Manuscript Parchment Leaves 8, 26, 47, and 68, our reference AA84050, 49, 48, 47 respectively" (letter to Christine McCarthy, Chief Conservator for Special Collections, Yale University Library, dated February 17, 2009), p. 9, Table 1.

4. Solvent extraction consisted of treating the samples with hexane, ethanol, methanol, and water in order to remove lipids and oils. It was assumed that any lipids and oils were human residues from handling of the parchment and were not part of the original animal hides.

5. Hodgins (2009), pp. 5–7.

6. Most references to the folios write the folio number immediately after the f, e.g., f72r. This author created an Excel database of the properties of the individual folios, and inserted a preceding zero or zeros, e.g., f006v, f072r, in order to make the database sortable in Excel.

7. Panofsky, Erwin (1954). Letter to William F. Friedman, 19 March 1954.

8. Zandbergen, René (2020). *History of Research of the Voynich*.

9. D'Imperio, Mary E. (1978), p. 7.

10. Ibid., p. 34.

11. Ibid., p. 7.

12. Ibid., p. 43.

13. Ibid., p. 7.

14. Ibid., pp. 7, 8.

15. Ibid., p. 7.

16. Ibid., p. 7.

17. Ibid., p. 37.

18. Propositions other than those sourced from D'Imperio (1978) are from Zandbergen (2020).

19. Hannig, Rainer (2020). *Voynisch-Hebräisch: Der Weg zur Entzifferung von Rainer Hannig* [Voynich-Hebrew: The way to decipher by Rainer Hannig].

20. Stolfi, J. (1998). "An Interlinear Archive of Voynich Manuscript Transcriptions In EVA," December 5, 1998.

21. Saxe, John Godfrey (1881). "The Blind Men and the Elephant" in *The Poems of John Godfrey Saxe* (Boston: Houghton, Mifflin and Company), p. 260.

22. This author will admit to buying a postcard of "The Face on Mars," probably at a seaside shop in Dubai, as well as a copy of *Unusual Martian Surface Features* by Vincent DiPietro and Gregory Molenaar, and to watching the movie *Mission to Mars* (2000).

Chapter 3

1. From Bob Dylan, "A Hard Rain's Gonna Fall" © Sony/ATV Music Publishing LLC.

2. Author's count based on import of the v101 transliteration into Microsoft Excel.

3. Nom de plume of Timothy Rayhel (1957–2014).

4. Author's count based on import of the v101 transliteration into Microsoft Word as unformatted text, treating all spaces, uncertain spaces, and line breaks as delimiters of "words," and treating all strings denoting "rare glyphs" and "unidentifiable glyphs" as single glyphs.

5. Norvig, Peter (2013?). English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU.

6. Ibid.

7. By this author's count, in the whole manuscript there are just 324 cases where a "gallows" glyph is at the end of a glyph string (the term "end" being understood to mean the rightmost glyph).

8. In Latin, "er" is the most common digram, with a frequency of 2.40 percent, followed by "um" with a frequency of 2.19 percent. Source: Stefan Trost Media, Syllable Frequency: Latin (Latina) https://www.sttmedia.com/syllablefrequency-latin.

9. Cappelli, Adriano (1929). Manuali Hoepli: *Lexicon Abbreviaturarum: Dizionario di Abbreviatture Latine ed Italiane* [Hoepli Manuals: Lexicon of Abbreviation: Dictionary of Latin and Italian Abbreviations], 3rd edition, revised and corrected (Milan: Ulrico Hoepli), table IV, between pages LX and LXI. Translation from Italian by author.

10. Ibid., p. LX.

11. Ibid., p. LXI.

12. Gessmann, G. W. (1922) *Die Geheimsymbole Der Alchymie, Arzneikunde Und Astrologie Des Mittelalters* (Berlin, Verlag von Karl Siegismund), p. 5. The original reads: "… die Tausende und aber Tausende von zusammengesetzten symbolischen Darstellungen, welche im Mittelalter geschaffen wurden, zusammenzutragen, wäre die Arbeit eines Menschenlebens." Translation from German by author.

13. Voynich, Wilfred (1921). "A Preliminary Sketch of the History of the Roger Bacon Cipher Manuscript" (Philadelphia: Transactions of the College of Physicians of Philadelphia).

14. Gessmann, G. W. (1922), pp. 16–17.

15. D'Imperio, Mary E. (1976). *The Voynich Manuscript: An Elegant Enigma* (Fort George G. Meade, MD: National Security Agency), p. 69.

16. Gessmann, G. W. (1922), p. LXXV.

17. Ibid., p. CII.

18. Ibid., p. XXV.

19. Ibid., p. CXX.

20. Poe, Edgar Allan. *The Works of Edgar Allan Poe, Vol. 1*.

21. Doyle, Sir Arthur Conan, *The Return of Sherlock Holmes*: Chapter 3, "The Adventure of the Dancing Men."

Chapter 4

1. Source: Bayart, Frédéric. Cryptographie et codes secrets, The Beale Papers [digital file].

2. Source: Hutchins, E. R. (ed.) (1887). *Biennial report of the Bureau of Labor Statistics for the State of Iowa*, 2nd ed., 1886–87, p. 10.

3. Zipf, George Kingsley (1929). "Relative Frequency as A Determinant of Phonetic Change" in *Harvard Studies in Classical Philology*, 40, 1–95. doi:10.2307/310585 [digital file].

4. Zipf, George Kingsley (1932). *Selected Studies of the Principle of Relative Frequency in Language* (Cambridge, MA: Harvard University Press) [digital file].

5. Zipf, George Kingsley (1929), pp. 3, 4.

6. Source: Sketch Engine, *Brown corpus: Corpus of American English* [digital file].

7. Source: Rugg, Gordon and Gavin Taylor (2016). *The Rugg and Taylor "Cryptologia" article on the Voynich Manuscript*, September 25, 2016 [digital file].

8. Ibid.

9. Ibid.

10. Ibid.

11. The function was =1+527*RAND(), rounded to the nearest integer.

Chapter 5

1. Tiltman, John H. (1976): Foreword, in D'Imperio, Mary E., *The Voynich Manuscript: An Elegant Enigma* (National Security Agency/Central Security Service: Fort George G. Meade, MD), p.

vii.

2. The mostly widely spoken modern Romance languages are Catalan, French, Galician, Portuguese, Romanian, and Spanish.

3. Zandbergen, René, Papers on the Voynich Manuscript, August 1, 2004; Station Hypo, "Remembering Capt. Prescott H. Currier, USN," *Cryptologist*, June 4, 2019.

4. D'Imperio, Mary E. (ed.) (1976). *New Research on the Voynich Manuscript: Proceedings of a seminar, November 30, 1976* (National Security Agency, Fort George G. Meade, MD), p. 20.

5. Ibid., p. 21.

6. The distance-decay theory states, in essence, that human interactions (including trade, investment, tourism, correspondence, marriage, and friendship) become less probable with increasing distance between the humans concerned. See for example Lilian S.C. Pun-Cheng, Distance decay, in *The International Encyclopedia of Geography*.

7. The languages of Albania, Bulgaria, and the former Yugoslavia.

8. The most widely spoken modern Slavic languages are Belarussian, Bulgarian, Czech, Polish, Russian, Serbo-Croat, Slovak, Slovenian, and Ukrainian.

9. Author's count from Microsoft Word, excluding punctuation.

10. D'Imperio, Mary E. (ed.) (1976), p. 22.

11. Ibid., p. 23.

12. Ibid., p. 23.

13. The Congress of the United States, Congress, *The unanimous Declaration of the thirteen united States of America*, July 4, 1776.

14. Shakespeare, William (1599). *Julius Caesar*, act 3, scene 2, line 73.

15. Ibid., p. 24.

16. Ibid., p. 25.

17. Bruns, Dave and Lisa Bruns, Exceljet: Count cells equal to case sensitive, https://exceljet.net/formula/count-cells-equal-to-case-sensitive.

18. Shelley, Percy Bysshe, "Ozymandias" in *Miscellaneous and Posthumous Poems of Percy Bysshe Shelley* (London: W. Benbow, 1826), 100. Additional letters generated by RAND() function in Excel.

19. Author's calculation.

20. Author's calculation.

21. Or in some cases the first fifty, if the total number of such words was less than 300.

22. Bulgarian, Danish, Icelandic, Macedonian, Norwegian (Bokmål), Norwegian (Nynorsk), Romanian, Swedish.

23. Author's calculation.

Chapter 6

1. D'Imperio, Mary E. (1978): *The Voynich Manuscript: An Elegant Enigma* (National Security Agency, Central Security Service, Fort George G. Meade, MD).

2. Ibid., p. ix.

3. Structural Linguistics, in this author's understanding, is the study of languages viewed not as collections of words but as structures with inter-related (even interlocking) parts, in other words, languages as analogous to machines.

4. D'Imperio, Mary E. (undated, fall 1978?): An Application of PTAH to the Voynich Manuscript (U) (National Security Agency, Fort George G. Meade, MD, DOCID: 3562200).

5. D'Imperio, Mary E. (June 1978): An Application of Cluster Analysis to the Question of "Hands" and "Languages" · in the Voynich Manuscript (National Security Agency, Fort George G. Meade, MD, DOCID: 6588666).

6. D'Imperio, Mary E. (undated, ca.1978): An Application of Cluster Analysis and Multidimensional Scaling to the Question of "Hands" and "Languages" in the Voynich Manuscript (National Security Agency, Fort George G. Meade, MD, DOCID: 636902).

7. D'Imperio, Mary E. (1978).

8. Ibid, pp. 6, 11.

9. Ibid, p. 11.

10. Ibid, p. 23.

11. Sravana Reddy and Kevin Knight, "What We Know About the Voynich Manuscript," p.79, in Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, pp. 78–86, Portland, Oregon, June 24, 2011.

12. Wolfram Alpha Computational Intelligence, *A Midsummer Night's Dream*.

13. Ibid.

14. Ibid.

15. D'Imperio, Mary E. (Fall 1978?)

16. Ambinder, Marc (2015). "The return of an intelligence code word with a storied history," in *The Week*, January 9, 2015.

17. D'Imperio, Mary E. (Fall 1978?), pp. 72, 74

18. Ibid, p. 76.

19. Ibid, p. 78.

20. Ibid, p.82.

21. Ibid, p. 83. This author capitalized the parsed string and the parsing phrases for clarity.

22. Ibid, p. 86.

23. Ibid, pp. 89–90.

Chapter 7

1. A. A. Markov, "An Example of Statistical Analysis of the Text of *Eugene Onegin* Illustrating the Association of Trials into a Chain," *Bulletin de l'Academie Imperiale des Sciences de St. Petersburg*, ser. 6, vol. 7 (1913), pp. 153–162.

2. Grinstead, Charles M., and J. Laurie Snell, *Introduction to Probability*, p. 465.

3. Stefan Trost Media, Syllable Frequency: English, https://www.sttmedia.com/syllablefrequency-english.

4. D'Imperio, Mary E. (1978?). "An Application of PTAH to the Voynich Manuscript (U)," (Fort George G. Meade, MD: National Security Agency), p. 66.

5. Ibid, p. 76.

6. Reddy, Sravana and Kevin Knight (2011). "What We Know About the Voynich Manuscript," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 78–86, Portland, Oregon, 24 June 2011.

7. The Buckwalter transliteration system uses a unique one-to-one mapping between Arabic characters (including diacritics) and modern keyboard symbols.

8. Sravana and Knight (2011), p. 80.

9. Ibid, p. 83.

10. Ibid, p. 84.

11. Montemurro, M. A., Zanette, D. H. (2013). "Keywords and Co-Occurrence Patterns in the Voynich Manuscript: An Information-Theoretic Analysis." PLoS ONE 8(6): e66344. doi:10.1371/journal.pone.0066344.

12. Stolfi, Jorge (1998). Reeds/Landini's interlinear file in EVA, version 1.6e6, December 28, 1998.

13. Author's count from Microsoft Word, based on Augustine: *Confessions*.

14. Hardy, Grant (1994). "Can an Ancient Chinese Historian Contribute to Modern Western Theory? The Multiple Narratives of Ssu-Ma Ch'ien" in *History and Theory*, February 1994, Vol. 33, No. 1, p. 22.

15. National Center for Biotechnology Information (NCBI) (2021). Sample GenBank Record, (Bethesda, MD: NCBI), as updated on January 12, 2021.

16. Montemurro and Zanette (2013), p. 6.

17. Author's count from Microsoft Word, based on unformatted text imported from v101 transliteration, with uncertain spaces treated as spaces.

18. Ibid., p. 2.

19. Ibid., p. 4.

20. Hauer, Bradley and Grzegorz Kondrak (2016). "Decoding Anagrammed Texts Written in an Unknown Language and Script," in *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 75–86, April 2016.

21. Ibid., p. 84.

22. Author's count from Microsoft Word, based on Lex Mercatoria, "United Nations Universal Declaration of Human Rights 1948," excluding spaces and punctuation.

23. For this purpose, we used Stefan Trost Media, Alphabet and Character Frequency: English, https://www.sttmedia.com/characterfrequency-english.

24. Busse, Ulrich (2002). Linguistic Variation in the Shakespeare Corpus: Morpho-syntactic Variability of second person pronouns (Amsterdam: John Benjamins Publishing, 29 November 2002), p. 1.

25. Author's count from Microsoft Word.

26. Istituto Opera del Vocabolario Italiano: Corpus OVI – Guida ai contenuti.

27. Author's interpretation of page f001r, line 1.

28. Ibid.

29. Zelinka, Ivan, Oldrich Zmeskal, Leah Windsor, Zhiqiang Cai (2019), "Unconventional Methods In Voynich Manuscript Analysis" in *Mendel Soft Computing Journal*, Volume 25, No.1 (Brno, Czech Republic: Mendel, June 2019).

30. Author's count from Microsoft Word, based on Hemingway, Ernest (1952). *The Old Man and the Sea* (New York: Charles Scribner's Sons).

31. Author's count from Microsoft Word, based on Zandbergen, Rene: C-D The original transliteration of D'Imperio and Currier.

32. Zelinka et al (2019), p. 10.

33. Ibid., p. 13.

34. Layfield, Colin, Lonneke van der Plas, Michael Rosner, John Abela (2020). "Word Probability Findings in the Voynich Manuscript," in *Proceedings of 1st Workshop on Language Technologies for*

Historical and Ancient Languages (Language Resources and Evaluation Conference (LREC 2020), Marseille, May 11–16, 2020), pp. 74–78.

35. Ibid., pp. 76, 77.

36. D'Imperio, Mary E. (1976). *New Research on the Voynich Manuscript: Proceedings of a seminar* (Fort George G. Meade, MD: National Security Agency, 30 November), p. 51.

Chapter 8

1. For the moment, we are ignoring the pages in languages that Currier did not identify, which contain only 16,025 glyphs.

2. The correlation coefficient is expressed as R squared, calculated by the RSQ function in Excel.

3. Moreno, Marsha Lynn (2005). "Frequency Analysis in Light of Language Innovation: Exploring letter frequencies across time, from the days of Old English to the days of now." (San Diego: University of California at San Diego, Spring 2005).

4. At this point the reader may be thinking why not dispense completely with the v101 alphabet, which is neither intuitive (in the sense of linking glyphs with similarly shaped letters, except in a few cases), nor efficient (in the sense of linking common glyphs with common letters). This is an entirely reasonable proposition. The v101 keyboard assignments used the standard American keyboard with its Latin letters, Arabic numerals, and punctuation symbols. Today we have Unicode, which allows transliterations and substitutions on an industrial scale.

5. Author's calculations. For details see appendix 7.

6. Wolfram Alpha, https://www.wolframalpha.com/input/?i=macbeth.

7. Author's calculation.

8. Author's counts based on texts from Project Gutenberg.

9. Wolfram Alpha, https://www.wolframalpha.com/input/?i=A+Midsummer+Night%27s+Dream.

10. Author's count based on *La Divina Commedia*.

11. Author's count based on *Partenopeus de Blois*, ca. 1180.

Chapter 9

1. D'Imperio, Mary E. (1978). *The Voynich Manuscript: An Elegant Enigma*, p. 8.

2. Friedman, Elizebeth (1962). "The Most Mysterious MS. Still an Enigma." *Washington Post*, 5 August 1962, quoted in D'Imperio, Mary E. (1978). The Voynich Manuscript: An Elegant Enigma, p. 8. This author has not located the original document.

3. Petersen. Theodore C. (1966). *Hand Transcript and Concordance of the Vov'nich Manuscript and Other Working Papers*. In the Friedman Collection, George Marshall Library, Lexington, Virginia. Quoted in D'Imperio, Mary E. (1978). *The Voynich Manuscript: An Elegant Enigma*, p. 8.

4. Zandbergen, Rene (2016). Pre-1960s tentative herb identifications. http://www.voynich.nu/extra/herb_oldid.html, December 3, 2016.

5. Sherwood, Edith, PhD and Erica Sherwood (2015). The Voynich Botanical Plants. http://www.edithsherwood.com/voynich_botanical_plants/index.php.

6. Ibid, p. 4.

7. US National Plant Germplasm System, Species of Cannabis, https://npgsweb.ars-grin.gov/gringlobal/taxonomylist?category=species&type=genus&value=Cannabis&id=2034.

8. Melbourne University, Multilingual Multiscript Plant Name Database, http://www.plantnames.unimelb.edu.au/new/Sorting/Admin.html.

9. Monteblanco. Luis Del Pino, S.L., http://monteblanco.bio/shop/extra-virgin-olive-oil-picual/.

10. Global Biodiversity Information Facility, Botrychium lunaria (L.) Sw. https://www.gbif.org/species/2650165.

11. Personal correspondence with VR (Voynich Researcher, name withheld).

12. Ibid.

13. Ibid.

14. Monteiro De Carvalho, José (1765). *Diccionario Portuguez Das Plantas, Arbustos etc*. (Lisbon), pp. 563, 564 (translation by author).

15. with apologies to Sherwood, Edith, PhD (2015), The Voynich Manuscript Decioded? Part II, http://www.edithsherwood.com/voynich_decoded_part2/page2.php.

16. For the mathematically minded, the formula is =1+25*RAND().

Chapter 10

1. International Astronomical Union, List of IAU-approved Star Names as of January 1, 2021.

2. de Lacaille, Nicolas-Louis (1757). *Astronomiae fundamenta novissimis solis et stellarum observationibus stabilita, Lutetiae in Collegio mazarineo et in Africa ad caput Bonae Spei peractis a Nicolao Ludovico de La Caille*. (Paris: J.-J. J. Stephani Collombat), pp. 227, 233.

3. Toomer, Gerald J., *Ptolemy's ALMAGEST Translated and Annotated by G. J. Toomer* (London: Duckworth, 1984), p. 361.

4. Verbunt, F. and R. H. van Gent, "The star catalogues of Ptolemaios and Ulugh Beg, Machine-readable versions and comparison with the modern HIPPARCOS Catalogue," in *Astronomy & Astrophysics*, 544, A31 (2012), p. 3.

5. Allen, Richard Hinckley (1899). *Star Names, Their Lore and Meaning* (New York, Leipzig, London, and Paris: G. E. Stechert), pp. 78–83.

6. Ibid., pp. 383–386.

7. Ibid., pp. 391–403.

Chapter 11

1. Cappelli, Adriano (1929). *Manuali Hoepli: Lexicon Abbreviaturarum: Dizionario di Abbreviatture Latine ed Italiane* [Hoepli Manuals: Lexicon of Abbreviation: Dictionary of Latin and Italian Abbreviations], 3rd Edition, Revised and Corrected, (Milano: Ulrico Hoepli).

2. Ibid., p. XII.

3. McGillivray, Barbara and Kilgarriff, Adam (2013). "Tools for historical corpus research, and a corpus of Latin," in Paul Bennett, Martin Durrell, Silke Scheible, Richard J. Whitt (eds.), New Methods in Historical Corpus Linguistics. (Tübingen: Narr).

4. Author's calculations based on The Packard Humanities Institute, PHI Latin Texts.

5. The average "word" is longer in v104 because some v101 glyphs are disaggregated into more basic shapes.

6. Ibid.

7. Internet Movie Data Base, *Gladiator*.

8. Translation by author.

9. Author's research based on The Packard Humanities Institute, PHI Latin Texts.

10. Ibid.

Chapter 12

1. For this purpose, we used the 1,000 most frequent words in the medieval sub corpus of the LatinISE corpus.

2. Author's calculation based on the reference corpus.

3. Shakespeare, William: *Macbeth* (act 5, scene 5, lines 19–22).

4. Reddit user u/Marc_Op [name withheld], "triple reduplication."

5. Author's translation.

6. Dylan, Bob: *Corrina, Corrina* lyrics © Special Rider Music, Universal Tunes.

7. Dylan, Bob: *Visions of Johanna* lyrics © Sony/ATV Music Publishing LLC.

8. Bagley, C.P. (1966): "Cantigas De Amigo and Cantigas De Amor," *Bulletin of Hispanic Studies*, 43:4, 241-252, DOI: 10.1080/14753826620000343241.

9. Lopes, Graça Videira; Ferreira, Manuel Pedro et al. (2011–), Cantigas Medievais Galego Portuguesas [online database], Lisboa: Instituto de Estudos Medievais, FCSH/NOVA. [Information retrieved on 20 August 2021].

Chapter 13

1. Neal, Philip (undated). The letter of Athanasius Kircher to Theodorus Moretus (1639) [Latin original].

2. Author's translation.

3. Cf. Findlen, Paula (ed.), 2004: *Athanasius Kircher: The Last Man Who Knew Everything*.

4. Cf. Wilkes, John J. (1995), *The Illyrians* (Oxford: Blackwell).

5. A stade was 150 to 200 meters; so 6,000 stades would be 900 to 1,200 kilometers (560 to 745 statute miles). A Roman mile was about 0.92 modern statute miles, so 750 Roman miles would be about 690 statute miles.

6. Appianus Alexandrinus, *Illyrike* (date unknown), p1, quoted in Wilkes, John J. (1995), *The Illyrians*, Oxford: Blackwell, p. 91

7. Unknown author (ca. 350 BC), *Periplus or Coastal Passage*, quoted in Wilkes, John J. (1995), The Illyrians, Oxford: Blackwell, pp. 94–96.

8. Kircher, Athanasius (1679): *Turris Babel sive Archontologia*, Book III, chapter 6, p. 215 (Amsterdam: Janssonio-Waesbergiana, 1679); translation by author.

9. Shepherd, William R., *Historical Atlas* (1911), pp. 42–43, The Roman Empire about 395.

10. Kircher, Athanasius (1679); chapter 1, p. 131; translation by author.

11. Ibid., chapter 6, pp. 215–216; translation by author.

12. Ibid., chapter 6, p. 216; translation by author.

13. Russell, J. C. (1958). "Late ancient and medieval population." Philadelphia, PA: American Philosophical Society.

14. Peter Turchin, Jonathan M. Adams, & Thomas D. Hall (2006): "East-West Orientation of Historical Empires and Modern States," in *Journal of World-Systems Research*, xii, December 2006, p. 222.

15. de Vries, Jan, "Population" in Thomas A. Brady, Heiko A. Oberman and James D. Tracy, *Handbook of European History, 1400–1600* (Leiden: Brill, 1994), p. 13.

16. Neal, Philip (undated). The letter of Athanasius Kircher to Theodorus Moretus (1639) [English translation].

17. Klein, Jared; Brian, Joseph; Fritz, Matthias (2018). *Handbook of Comparative and Historical Indo-European Linguistics*. Walter de Gruyter. p. 1800.

18. Author's counts.

19. Dr. Michiel de Vaan, personal communication, November 8, 2021

20. Ibid.

21. Ibid.

Chapter 14

1. Davis, Lisa Fagin (2019). "How Many Glyphs and How Many Scribes? Digital Paleography and the Voynich Manuscript." *Manuscript Studies* Vol. 5, issue 1, article 6 (Spring 2020). https://repository.upenn.edu/mss_sims/vol5/iss1/6

2. Bowern, Claire L., and Daniel E. Gaskell (2022). "Enciphered after all? Word-level text metrics are compatible with some types of encipherment." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper6.pdf

3. Hermes, Jürgen (2022). "Polygraphia III: The cipher that pretends to be an artificial language." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceurws.org/Vol-3313/paper7.pdf

4. Dunin, Elonka, and Klaus Schmeh (2022). "The Voynich Manuscript Compared with Other Encrypted Books." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper15.pdf

5. Boxer, Alexander (2022). "Fingerprinting Gibberish: A Quantitative Comparison of the Voynich and Sloane MS 3188." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper1.pdf

6. British Library, Digitised Manuscripts, *Sloane MS 3188*. https://www.bl.uk/manuscripts/FullDisplay.aspx?ref=Sloane_MS_3188

7. Laycock, Donald C. (2001). *The Complete Enochian Dictionary*. Weiser Books, San Francisco, 2001, pp. 33–5.

8. Whitby, Christopher Lionel (1988). *John Dee's Actions with Spirits: 22 December 1581 to 23 May 1583*, Volume 2. Garland, New York, 1988.

9. Gaskell, Daniel E., and Claire L Bowern (2022). "Gibberish after all? Voynichese is statistically similar to human produced samples of meaningless text." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper4.pdf.

10. Caruana, Andrew, Colin Layfield, and John Abela (2022). "An Analysis of the Relationship between Words within the Voynich Manuscript." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper8.pdf.

11. Lindemann, Luke (2022). "Crux of the MATTR: Voynichese Morphological Complexity." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceurws.org/Vol-3313/paper9.pdf.

12. Stafford, Tavi (2022). "Seven Habits of Highly Eccentric Paragraphs." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper13.pdf.

13. Author's analysis based on Gutenberg text. https://www.gutenberg.org/ebooks/1012.

14. Feaster, Patrick (2022). "Rightward and Downward Grapheme Distributions in the Voynich Manuscript." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper12.pdf.

15. Painter, Katie, and Claire Bowern (2022). “Examining the history of Voynich glyphs using phylogenetic methods.” In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper11.pdf.

16. Zattera, Massimiliano (2022). “A new Transliteration Alphabet brings new Evidence of Word Structure and Multiple “languages” in the Voynich Manuscript.” In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper10.pdf.

Afterword

1. Except in the case of the *Murder on the Orient Express* (to avoid spoilers, we shall say no more).

2. Norman, Jeremy (undated). “Relative Costs in Corbie of the Components of a Medieval Manuscript.”

3. Bozzolo, Carla and Ezio Ornato (1980). *Pour une histoire du livre manuscrit au moyen Age: Trois essais de codicologie quantitative* (Paris: Centre national de la recherche scientifique), p. 25.

4. Biblioteca Statale di Cremona nr. 87, colophon 14210. By courtesy of Dr. Aidan Conti, personal communication.

5. *Missal of Nicholas Lytlington* (1383–84), colophon 17993 (London, Westminster Abbey Library). By courtesy of Dr. Aidan Conti, personal communication.

6. Nouvelles acquisitions latines nr. 2288 f. 201v, nr. 2289 f. 259v, colophon 454 (Paris: Bibliothèque Nationale). By courtesy of Dr. Aidan Conti, personal communication.

7. Overty, Joanne Filippone (2008). “The Cost of Doing Scribal Business: Prices of Manuscript Books in England, 1300–1483” in *Book History*, Vol. 11, pp. 1–32 (Johns Hopkins University Press).

8. Gies, Joseph and Frances Gies (1981). *Life in a Medieval City* (New York: HarperPerennial).

9. Phelps Brown, E. H., and Sheila V. Hopkins (1955). “Seven Centuries of Building Wages,” in *Economica*, New Series, Vol. 22, No. 87 (August 1955), pp. 195–206.

SOURCES

1. Books

Allen, Richard Hinckley. *Star Names, Their Lore and Meaning*. New York, Leipzig, London, and Paris: G. E. Stechert, 1899.

Appianus Alexandrinus. *Illyrike*. Alexandria, Egypt: unknown, ca. 100 BC.

Avakov, Alexander V. *Two Thousand Years of Economic Statistics, Volume 1*. New York: Algora, December 14, 2009.

Brady, Thomas A.; Heiko A. Oberman; James D. Tracy. *Handbook of European History, 1400–1600*. Leiden: Brill, 1994.

Busse, Ulrich. *Linguistic Variation in the Shakespeare Corpus: Morphosyntactic Variability of second person pronouns*. Amsterdam: John Benjamins, November 29, 2002.

Clemens, Raymond (ed.). *The Voynich Manuscript*. New Haven, CT: Yale University Press, November 1, 2016.

de Lacaille, Nicolas-Louis. *Astronomiae fundamenta novissimis solis et stellarum observationibus stabilita, Lutetiae in Collegio mazarineo et in Africa ad caput Bonae Spei peractis a Nicolao Ludovico de La Caille*. Paris: J. J. Stephani Collombat, 1757.

Frucht, Richard C., (ed.). *Eastern Europe*. Santa Barbara, CA: ABC-CLIO, 2005.

Gies, Joseph, and Frances Gies. *Life in a Medieval City*. New York: HarperPerennial, 1981.

Hemingway, Ernest. *The Old Man and the Sea*. New York: Charles Scribner's Sons, 1952.

Hutchins, E. R. (ed.). *2nd Biennial report of the Bureau of Labor Statistics for the State of Iowa*. Des Moines, IA: Iowa BLS, 1887.

Klein, Jared; Joseph Brian; Matthias Fritz. *Handbook of Comparative and Historical Indo-European Linguistics*. Berlin: Walter de Gruyter, 2018.

Petersen, Theodore C. *Hand Transcript and Concordance of the Voynich Manuscript and Other Working Papers*. Lexington, VA: manuscript, 1966.

Shelley, Percy Bysshe. *Miscellaneous and Posthumous Poems of Percy Bysshe Shelley*. London: W. Benbow, 1826.

Toomer, Gerald J. *Ptolemy's ALMAGEST Translated and Annotated by G. J. Toomer*. London: Duckworth, 1984.

unknown. *Periplus or Coastal Passage*. Unknown: manuscript, ca. 350–300 BC.

Winter, Jay. *The Complete Voynich Manuscript: Digitally Enhanced Researchers Edition*. Morrisville, NC: Lulu, October 17, 2015.

2.1 Digital files: e-books

Augustine of Hippo. *Confessiones*. Hippo, Numidia: manuscript, 397 to 400. https://www.gutenberg.org/cache/epub/33849/pg33849.txt.

Buzuku, Gjon. *Meshari (Missale) (1555), TITUS version by Jost Gippert*. Frankfurt: University of Frankfurt, 09-Mar-2003. https://titus.uni-frankfurt.de/texte/etcs/alban/buzuku/buzukt.htm.

Cappelli, Adriano. *Manuali Hoepli: Lexicon Abbreviaturarum: Dizionario di Abbreviatture Latine ed Italiane*. Milan: Ulrico Hoepli, 1929. https://archive.org/details/CappelliDizionarioDiAbbreviature.

Crapelet, Georges Adrien. *Anciens monuments de l'histoire et de la langue françaises. 12, Partonopeus de Blois. Tome premier-tome second / publ. pour la première fois, d'après le ms. de la Bibliothèque de l'Arsenal, avec trois fac-sim*. Paris: Imprimerie de Crapelet, 1834. https://gallica.bnf.fr/ark:/12148/bpt6k4120k.

Currier, Prescott H. "Some Important New Statistical findings" in *New Research on the Voynich Manuscript: Proceedings of a seminar*, edited by Mary E. D'Imperio, 20–26. Fort George G. Meade, MD: National Security Agency, November 30, 1976. https://www.nsa.gov/Portals/70/documents/news-features/declassified-documents/voynich/proceedings-of-a-seminar-30-november-1976.pdf.

Currier, Prescott H. "The Voynich Manuscript, some notes and observations," in *New Research on the Voynich Manuscript: Proceedings of a seminar*, edited by Mary E. D'Imperio, 61–73. Fort George G. Meade, MD: National Security Agency, November 30, 1976. https://www.nsa.gov/Portals/70/documents/news-features/declassified-documents/voynich/proceedings-of-a-seminar-30-november-1976.pdf.

Dante Alighieri. *La Divina Commedia*. Foligno, Italy: Johann Numeister and Evangelista Angelini da Trevi, ca. 1308–1320. https://www.gutenberg.org/ebooks/1000.

D'Imperio, Mary E. *The Voynich Manuscript: An elegant enigma*. Fort George G. Meade, MD: National Security Agency, 1978. https://apps.dtic.mil/dtic/tr/fulltext/u2/a070618.pdf.

D'Imperio, Mary E. *The Voynich Manuscript: An elegant enigma*. Fort George G. Meade, MD: National Security Agency, 1978. https://www.nsa.gov/Portals/70/documents/about/cryptologic-heritage/historical-figures-publications/publications/pre-modern/voynich_manuscript.pdf?ver=2019-08-09-093405-303.

D'Imperio, Mary E. (ed.) *New Research on the Voynich Manuscript: Proceedings of a seminar*. Fort George G. Meade, MD: National Security Agency, November 30, 1976. https://www.nsa.gov/Portals/70/documents/news-features/declassified-documents/voynich/proceedings-of-a-seminar-30-november-1976.pdf.

Doyle, Sir Arthur Conan. *The Return of Sherlock Holmes*: Chapter 3: The Adventure of the Dancing Men. New York: McClure, Phillips, 1905. https://en.wikisource.org/wiki/The_Return_of_Sherlock_Holmes/Chapter_3.

Gessmann, G. W. *Die Geheimsymbole Der Alchymie, Arzneikunde Und Astrologie Des Mittelalters*. Berlin: Siegismund, 1922. https://ia601006.us.archive.org/7/items/DieGeheimsymboleDerAlchymie/gessmann-g-geheimsymbole-1922-RTL010443-LowRes.pdf.

Grinstead, Charles Miller, and J. Laurie Snell. *Introduction to Probability*. Providence, RI: American Mathematical Society, 1997. https://archive.org/details/flooved3489/.

Kircher, Athanasius. *Turris Babel, sive Archontologia*. Amsterdam: Janssonio-Waesbergiana, 1679. http://www.archive.org/details/turrisbabelsivea00kirc.

Poe, Edgar Allen. *The Works of Edgar Allan Poe, Volume 1*. New York P. F Collier & Son, 1903. https://gutenberg.org/ebooks/2147

Pun-Cheng, Lilian S.C. "Distance decay," in *The International Encyclopedia of Geography*, edited by Douglas Richardson et al. New York, NY: John Wiley & Sons, Ltd., 2017. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118786352.wbieg0179.

Sommerhoff, Johann Christof. *Lexicon pharmaceutico-chymicum latino-germanicum et germanico-latinum*. Nuremberg, Germany: Froberg for Zieger and Lehmann, 1701. https://archive.org/details/bub_gb_1pRDAAAAcAAJ.

Tiltman, John H. *The Voynich manuscript: The most mysterious manuscript in the world*. Fort George G. Meade, MD: National Security Agency, 1967. https://web.archive.org/web/20111018025101/https://www.nsa.gov/public_info/_files/tech_journals/Voynich_Manuscript_Mysterious.pdf.

unknown [possibly J. B. Ward]. *The Beale Papers, containing authentic statements regarding the Treasure Buried in 1819 And 1821, near Bufords, in Bedford County, Virginia, and Which Has Never Been Recovered*. Lynchburg, VA: Virginian Book & Job Print, 1885. http://www.bibmath.net/crypto/ancienne/bealetextes.pdf.

unknown. *Cipher Manuscript / Voynich Manuscript*. unknown: manuscript, ca. 1401–1599? https://collections.library.yale.edu/catalog/2002046.

Wilkes, John J. *The Illyrians*. Oxford: Blackwell, 1995. https://vk.com/doc278943367_472533214?hash=75a22063d7379528f6.

Wycliffe, John et al. *Holy Bible, New Testament*. England: manuscript, ca. 1395. https://ebible.org/pdf/engWycliffe/engWycliffe_nt.pdf.

Zipf, George. *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press, 1932. https://pure.mpg.de/rest/items/item_2407800/component/file_2459540/content.

2.2 Digital files: journal articles

Davis, Lisa Fagin. "How Many Glyphs and How Many Scribes? Digital Paleography and the Voynich Manuscript." *Manuscript Studies*, Vol. 5, issue 1, article 6 (Spring 2020). https://repository.upenn.edu/mss_sims/vol5/iss1/6.

Hauer, Bradley, and Grzegorz Kondrak. "Decoding Anagrammed Texts Written in an Unknown Language and Script." *Transactions of the Association for Computational Linguistics* 4 (April 2016), 75–86. https://aclanthology.org/Q16-1006.pdf.

Layfield, Colin, Lonneke van der Plas, Michael Rosner, John Abela. "Word Probability Findings in the Voynich Manuscript." *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages* (May 11–16, 2020), 74-78. https://aclanthology.org/2020.lt4hala-1.11.pdf.

Overty, Joanne Filippone. "The Cost of Doing Scribal Business: Prices of Manuscript Books in England, 1300-1483." *Book History* 11 (2008), 1–32. https://www.jstor.org/stable/30227411.

Phelps Brown, E. H., and Sheila V. Hopkins. "Seven Centuries of Building Wages." *Economica, n.s.* 22, no. 87 (August 1955), 195–206. https://www.jstor.org/stable/2626898.

Reddy, Sravana, and Kevin Knight. "What We Know About the Voynich Manuscript." *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (June 24, 2011), 78–86. https://aclanthology.org/W11-1511.pdf.

Voynich, Wilfrid M. "A Preliminary Sketch of the History of the Roger Bacon Cipher Manuscript." *Transactions of the College of Physicians of Philadelphia* 43 (April 20, 1921), 415–430. https://archive.org/details/transactionsofco3431coll/page/414/mode/2up?view=theater

Zelinka, Ivan, Oldrich Zmeskal, Leah Windsor, Zhiqiang Cai. "Unconventional Methods In Voynich Manuscript Analysis." *Mendel Soft Computing Journal* 25, no.1 (June 2019), 1–14. https://mendel-journal.org/index.php/mendel/article/view/72/99.

Zipf, George. "Relative Frequency as A Determinant of Phonetic Change." *Harvard Studies in Classical Philology* 40 (1929), 1–95.

https://pure.mpg.de/rest/items/item_2407799_2/component/file_2407798/content.

2.3 Digital files: news articles

Ambinder, Marc. "The return of an intelligence code word with a storied history." *The Week*, January 9, 2015. https://theweek.com/articles/445540/return-intelligence-code-word-storied-history.

2.4 Digital files: website content

Bayart, Frédéric. "Cryptographie et codes secrets: The Beale Papers." Accessed November 19, 2022. http://www.bibmath.net/crypto/ancienne/bealetextes.pdf.

Beinecke Rare Book & Manuscript Library. "Voynich Manuscript: A mysterious, undeciphered manuscript." Accessed November 19, 2022. https://beinecke.library.yale.edu/collections/highlights/voynich-manuscript.

Bowern, Claire L., and Daniel E. Gaskell. "Enciphered after all? Word-level text metrics are compatible with some types of encipherment." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper6.pdf.

Boxer, Alexander. "Fingerprinting Gibberish: A Quantitative Comparison of the Voynich and Sloane MS 3188." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper1.pdf.

Brewer, Keagan. "'I beg your grace that you suppress this chapter or else allow it to be written in secret letters': The emotions of encipherment in late-medieval gynaecology." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper3.pdf.

Bruns, Dave and Lisa Bruns. "Exceljet: Count cells equal to case sensitive." Accessed November 19, 2022. https://exceljet.net/formula/count-cells-equal-to-case-sensitive.

Caruana, Andrew, Colin Layfield, and John Abela. "An Analysis of the Relationship between Words within the Voynich Manuscript." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper8.pdf.

CLARIN. "Historical corpora." Accessed November 19, 2022. https://www.clarin.eu/resource-families/historical-corpora.

DataBasic. "WordCounter" Accessed November 19, 2022. https://databasic.io/en/wordcounter/.

D'Imperio, Mary E. "An Application of Cluster Analysis and Multidimensional Scaling to the Question of "Hands" and "Languages" in the Voynich Manuscript." ca. 1978; accessed November 19, 2022. https://www.nsa.gov/Portals/70/documents/news-features/declassified-documents/tech-journals/application-of-cluster-analysis.pdf.

D'Imperio, Mary E. "An Application of Cluster Analysis to the Question of "Hands" and "Languages" in the Voynich Manuscript." 01/06/1978; accessed November 19, 2022. https://www.nsa.gov/Portals/70/documents/news-features/declassified-documents/voynich/an-application-of-cluster-analysis.pdf.

D'Imperio, Mary E. "An Application of PTAH to the Voynich Manuscript (U)." ca. 1978; accessed November 19, 2022. https://www.nsa.gov/news-features/declassified-documents/tech-journals/assets/files/application-of-ptah.pdf.

Dunin, Elonka, and Klaus Schmeh. "The Voynich Manuscript Compared with Other Encrypted Books." In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper15.pdf.

Feaster, Patrick. “Rightward and Downward Grapheme Distributions in the Voynich Manuscript.” In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper12.pdf.

Gaskell, Daniel E., and Claire L Bowern. “Gibberish after all? Voynichese is statistically similar to human produced samples of meaningless text.” In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper4.pdf.

Gies, Joseph, and Frances Gies. “Life in a Medieval City.” 1981; accessed November 19, 2022. https://machaut.weebly.com/money-in-the-middle-ages.html.

Global Biodiversity Information Facility. “Botrychium lunaria (L.) Sw.” Accessed November 19, 2022. https://www.gbif.org/species/2650165.

Hannig, Rainer. “Voynisch-Hebräisch—Der Weg zur Entzifferung von Rainer Hannig.” June 7, 2020; accessed November 19, 2022. https://www.rainer-hannig.com/app/download/15835537224/Voynich_Hebr%C3%A4isch_von_Rainer+Hannig_7_6_20.pdf.

Hermes, Jürgen. “Polygraphia III: The cipher that pretends to be an artificial language.” In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper7.pdf.

Hodgins, Gregory W. L. “Dating Results for Voynich Manuscript Parchment Leaves 8, 26, 47, and 68, our reference AA84050, 49, 48, 47 respectively.” May 14, 2021; accessed November 19, 2022. https://proto57.wordpress.com/2021/05/14/the-long-awaited-voynich-radiocarbon-report/.

Instituto da Lingua Galega. “Corpus Xelmírez, Corpus lingüístico da Galicia medieval.” Accessed November 19, 2022. http://ilg.usc.gal/xelmirez.

International Astronomical Union. “List of IAU-approved Star Names as of January 1st, 2021.” January 1, 2021; accessed November 19, 2022. https://www.iau.org/public/themes/naming_stars/.

Internet Movie Data Base. “Gladiator.” Accessed November 19, 2022. https://www.imdb.com/title/tt0172495/.

Istituto Opera del Vocabolario Italiano. “Corpus OVI - Guida ai contenuti.” Accessed November 19, 2022. http://gattoweb.ovi.cnr.it/HelpCorpora/Infoovi.html.

Leipzig University, Corpora Collection. “Galician Wikipedia corpus glg_wikipedia_2021.” Accessed November 19, 2022. https://corpora.uni-leipzig.de/en?corpusId=glg_wikipedia_2021.

Lindemann, Luke. “Crux of the MATTR: Voynichese Morphological Complexity.” In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper9.pdf.

University of Oslo. “United Nations Universal Declaration of Human Rights 1948.” Accessed November 19, 2022. https://www.jus.uio.no/english/services/library/treaties/02/2-01/declaration-human-rights.xml.

Lopes, Graça Videira; Manuel Pedro Ferreira et al. “Cantigas Medievais Galego Portuguesas.” Accessed November 19, 2022. http://cantigas.fcsh.unl.pt.

Lyons, James. “Practical Cryptography: Letter Frequencies for Various Languages.” Accessed November 19, 2022. http://practicalcryptography.com/cryptanalysis/letter-frequencies-various-languages/.

MacKinnon, Colin. “The Naturalization Papers of Wilfrid Voynich.” Accessed November 19, 2022. http://www.colinmackinnon.com/attachments/The_Naturalization_Papers_of_Wilfrid_Voynich_W.pdf.

MacKinnon, Colin. “The Registration of Marriage of Wilfrid and Ethel Voynich.” Accessed November 19, 2022. http://www.colinmackinnon.com/attachments/The_Registration_of_Marriage_of_WMV_and_ELV_A.pdf.

MacKinnon, Colin. “The Travels of Ethel and Wilfrid Voynich and Anne M. Nill Between the U.S. and the U.K. According to American and British Records (55 KB).” Accessed November 19, 2022. http://www.colinmackinnon.com/attachments/The_Travels_of_Ethel_and_Wilfrid_Voynich_and_Anne_M._Nill_W.pdf.

MacKinnon, Colin. “The U.S.Bureau Of Investigation File on Wilfrid Voynich.” Accessed November 19, 2022. http://www.colinmackinnon.com/attachments/The_US_BOI_File_on_Voynich_p.pdf.

MacKinnon, Colin. “The Voynich “Coat of Arms” and Michael Voynich’s Flight from Siberia.” Accessed November 19, 2022. http://www.colinmackinnon.com/attachments/The_Voynich_Coat_of_Arms_and_Michael_Voynichs_Flight_from_Siberia_W.pdf.

MacKinnon, Colin. “The Voyniches in the 1901 Census of England and Wales.” Accessed November 19, 2022. http://www.colinmackinnon.com/attachments/The_Voyniches_in_the_1901_Census_of_England_and_Wales_W.pdf.

MacKinnon, Colin. “Wilfrid Voynich’s Certificate of Death and Grave Site.” Accessed November 19, 2022. http://www.colinmackinnon.com/attachments/Wilfrid_Michael_Voynich_Death_Certificate_A.pdf.

Mariño Paz, Prof. Ramón. “Edición Cantiga Ajuda, Ejercicios de Lengua y Literatura Gallega.” Accessed November 19, 2022. https://www.docsity.com/es/edicion-cantiga-ajuda/5673570/.

Melbourne University. “Multilingual Multiscript Plant Name Database.” Accessed November 19, 2022. http://www.plantnames.unimelb.edu.au/new/Sorting/Admin.html.

Moreno, Marsha Lynn. “Frequency Analysis in Light of Language Innovation: Exploring letter frequencies across time, from the days of Old English to the days of now.” 2005; accessed November 19, 2022. http://www.math.ucsd.edu/~crypto/Projects/MarshaMoreno/TimeComparisonFrequency.pdf.

MyLanguages. “Hebrew to Latin Converter.” Accessed November 19, 2022. http://mylanguages.org/hebrew_romanization.php

NASA. “Mars Global Surveyor image of the ‘Face.’” 2001; accessed November 19, 2022. https://mars.nasa.gov/mgs/msss/camera/images/4_6_face_release/face_alone.gif.

NASA. “Viking CD-ROM Volume 10, image F035A72.” July 25, 1976; accessed November 19, 2022. http://photojournal.jpl.nasa.gov/catalog/PIA01141.

National Center for Biotechnology Information (NCBI). “Sample GenBank Record.” Accessed November 19, 2022. https://www.ncbi.nlm.nih.gov/genbank/samplerecord/.

Neal, Philip. “Johannes Marcus Marci, Philosophia Vetus Restituta (1662) [English translation].” Accessed November 19, 2022. http://philipneal.net/voynichsources/marci_philosophia_english/.

Neal, Philip. “Notes on the letter of Georgius Barschius to Athanasius Kircher (1637).” Accessed November 19, 2022. http://philipneal.net/voynichsources/barschius_1637_notes/.

Neal, Philip. “Notes on the letter of Johannes Marcus Marci to Athanasius Kircher (1640).” Accessed November 19, 2022. http://philipneal.net/voynichsources/marci_1640_notes/.

Neal, Philip. “The letter of Athanasius Kircher to Theodorus Moretus (1639) [English translation].” Accessed November 19, 2022. http://philipneal.net/voynichsources/kircher_1639_english/.

Neal, Philip. “The letter of Georgius Barschius to Athanasius Kircher (1637) [English].” Accessed November 19, 2022. http://philipneal.net/voynichsources/barschius_1637_english/.

Neal, Philip. “The letter of Georgius Barschius to Athanasius Kircher (1637) [Latin].” Accessed November 19, 2022. http://philipneal.net/voynichsources/barschius_1637_latin/.

Neal, Philip. “The letter of Johannes Marcus Marci to Athanasius Kircher (1665) [English].” Accessed November 19, 2022. http://philipneal.net/voynichsources/marci_1665_english/.

Neal, Philip. “The letters of Johannes Marcus Marci to Athanasius Kircher [English].” Accessed November 19, 2022. http://philipneal.net/voynichsources/barschius_1637_english/.

Neal, Philip. “The letters of Johannes Marcus Marci to Athanasius Kircher [Latin].” Accessed November 19, 2022. http://philipneal.net/voynichsources/marci_kircher_latin/#2.

Neal, Philip. “The letters of Johannes Marcus Marci to Athanasius Kircher: Beinecke 408 [Latin].” Accessed November 19, 2022. http://philipneal.net/voynichsources/marci_kircher_latin/#beinecke.

Norman, Jeremy. “Relative Costs in Corbie of the Components of a Medieval Manuscript.” Accessed November 19, 2022. https://www.historyofinformation.com/detail.php?id=275.

Norvig, Peter. “English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU.” 2013? Accessed November 19, 2022. http://norvig.com/mayzner.html.

Packard Humanities Institute. “PHI Latin Texts.” 2013? Accessed November 19, 2022. https://latin.packhum.org/.

Painter, Katie, and Claire Bowern. “Examining the history of Voynich glyphs using phylogenetic methods.” In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper11.pdf

Panofsky, Erwin. “Letter to William F. Friedman, quoted by Jim Reeds.” March 19, 1954; accessed November 19, 2022. http://www.voynich.nu/q20/f116v_tr.txt

Rugg, Gordon, and Gavin Taylor. “The Rugg and Taylor “Cryptologia” article on the Voynich Manuscript.” September 25, 2016; accessed November 19, 2022. https://hydeandrugg.wordpress.com/2016/09/25/the-rugg-and-taylor-cryptologia-article-on-the-voynich-manuscript/

Rugg, Gordon. Replicating the Voynich Manuscript 13/04/2004; accessed November 19, 2022. http://web.archive.org/web/20060717030430/www.cs.keele.ac.uk/content/people/g.rugg/voynich/.

Shepherd, William R. “Historical Atlas, pp42-43, The Roman Empire about 395.” 1911; accessed November 19, 2022, https://commons.wikimedia.org/wiki/File:Roman_Empire_about_395.jpg.

Sherwood, Edith, PhD, and Erica Sherwood. “The Voynich Botanical Plants.” 2015; accessed November 19, 2022. http://www.edithsherwood.com/voynich_botanical_plants/index.php.

Sherwood, Edith, PhD. “The Voynich Botanical Plants 2.0.” 2013; accessed November 19, 2022, https://www.edithsherwood.com/voynich_botanical_plants_2/.

Sketch Engine. “Brown corpus: Corpus of American English.” Accessed November 19, 2022. https://www.sketchengine.eu/brown-corpus/.

Stafford, Tavi. “Seven Habits of Highly Eccentric Paragraphs.” In University of Malta, *International Conference on the Voynich Manuscript 2022*, November 30–December 1, 2022. https://ceur-ws.org/Vol-3313/paper13.pdf.

Station Hypo. “Remembering CAPT Prescott H. Currier, USN, Cryptologist.” June 4, 2019; accessed November 19, 2022. https://stationhypo.com/2019/06/04/remembering-capt-prescott-h-currier-usn-cryptologist/.

Stefan Trost Media. “Character Frequencies.” Accessed November 19, 2022. https://www.sttmedia.com/characterfrequencies.

Stefan Trost Media. “Character Frequency: Galician (Galego).” Accessed November 19, 2022. https://www.sttmedia.com/characterfrequency-galician.

Stolfi, Jorge S. “An Interlinear Archive of Voynich Manuscript Transcriptions in EVA.” December 5, 1998; accessed November 19, 2022. www.voynich.nu/data/beta/LSI_ivtff_0d.txt.

Stolfi, Jorge S. “Reeds/Landini’s interlinear file in EVA, version 1.6e6.” December 28, 1998; accessed November 19, 2022. https://www.ic.unicamp.br/~stolfi/voynich/98-12-28-interln16e6/.

Stolte, Daniel. “Experts determine age of book ‘nobody can read.’” February 10, 2011; accessed November 19, 2022. https://phys.org/news/2011-02-experts-age.html

Timm, Torsten. “How the Voynich Manuscript was created.” December 9, 2015; accessed November 19, 2022. https://arxiv.org/ftp/arxiv/papers/1407/1407.6639.pdf.

u?Marc Op (Reddit user). “Triple reduplication.” Accessed November 19, 2022. https://www.reddit.com/r/voynich/comments/9rj7nh/triple_reduplication/.

University of Oslo. “United Nations Universal Declaration of Human Rights 1948.” Accessed November 19, 2022. https://www.jus.uio.no/english/services/library/treaties/02/2-01/declaration-human-rights.xml.

US Defense Department. “Lieutenant Prescott Currier, USN.” September 17, 2020; accessed November 19, 2022, https://media.defense.gov/2020/Sep/17/2002499404/-1/-1/0/OTR%20GANG%20CURRIER.PDF.

US National Plant Germplasm System. “Genus Cannabis L.” Accessed November 19, 2022. https://npgsweb.ars-grin.gov/gringlobal/taxon/taxonomygenus?id=2034.

Wolfram Alpha Computational Intelligence. “A Midsummer Night’s Dream.” Accessed November 19, 2022. https://www.wolframalpha.com/input/?i=A+Midsummer+Night%27s+Dream.

Wolfram Alpha Computational Intelligence. “Macbeth.” Accessed November 19, 2022. https://www.wolframalpha.com/input/?i=Macbeth.

Yasumasa, Someya. “Do it Yourself Word Frequency Count.” Accessed November 19, 2022. http://www.someya-net.com/00-class09/wordCount.html.

Zandbergen, René. “Analysis Section (1/5) - Introduction.” January 2, 2019; accessed November 19, 2022. http://www.voynich.nu/a1_intro.html.

Zandbergen, René. “Analysis Section (2/5) - Character statistics.” December 22, 2020; accessed November 19, 2022. http://www.voynich.nu/a2_char.html.

Zandbergen, René. “Analysis Section (3/5) - Word structure.” March 25, 2019; accessed November 19, 2022. http://www.voynich.nu/a3_para.html.

Zandbergen, René. “Analysis Section (4/5) - Word statistics.” December 26, 2015; accessed November 19, 2022. http://www.voynich.nu/a4_word.html.

Zandbergen, René. “Analysis Section (5/5) - Sentences, paragraphs, sections.” January 2, 2019; accessed November 19, 2022. http://www.voynich.nu/a5_synt.html.

Zandbergen, René. “C-D The original transliteration of D’Imperio and Currier.” June 19, 2021; accessed November 19, 2022. http://www.ic.unicamp.br/~stolfi/voynich/mirror/gillogly/voynich.orig.

Zandbergen, René. “GC The v101 transliteration file.” June 19, 2021; accessed November 19, 2022. http://www.voynich.nu/data/voyn_101.txt.

Zandbergen, René. “History of research of the Voynich.” March 8, 2022; accessed November 19, 2022. www.voynich.nu/solvers.html.

Zandbergen, René. “Jacobus Horčický de Tepenec (Sinapius).” March 20, 2021; accessed November 19, 2022. http://www.voynich.nu/curricula.html#sinapius.

Zandbergen, René. “Last page.” December 4, 1998; accessed November 19, 2022. http://www.voynich.nu/q20/f116v_tr.txt.

Zandbergen, René. “Manuscripts Voynich acquired in 1911–1912.” September 18, 2020; accessed November 19, 2022. http://www.voynich.nu/extra/coll_rom_mss.html.

Zandbergen, René. “Papers on the Voynich Manuscript.” August 1, 2004; accessed November 19, 2022. http://www.voynich.nu/extra/curr_main.html.

Zandbergen, René. “The history of the Voynich MS.” October 10, 2018; accessed November 19, 2022. http://www.voynich.nu/history.html.

Zandbergen, René. “The origin of the Voynich MS.” July 26, 2017; accessed November 19, 2022. http://www.voynich.nu/origin.html.

Zandbergen, René. “The Radio-Carbon Dating of the Voynich MS.” December 5, 2020; accessed November 19, 2022. http://www.voynich.nu/extra/carbon.html.

Zandbergen, René. “The Voynich MS - Bibliography and references.” April 27, 2021; accessed November 19, 2022. http://www.voynich.nu/refs.html.

Zandbergen, René. “The Voynich MS - Biographies.” March 20, 2021; accessed November 19, 2022. http://www.voynich.nu/curricula.html#h13.

Zandbergen, René. “The Voynich MS - Picture Galleries.” September 29, 2016; accessed November 19, 2022. http://www.voynich.nu/gallery.html.

Zandbergen, René. “Voynich MS - 17th Century letters related to the MS.” March 23, 2019; accessed November 19, 2022. http://www.voynich.nu/letters.html.

3. Journal articles

Markov, Andrei Andreyevich. “An Example of Statistical Analysis of the Text of Eugene Onegin Illustrating the Association of Trials into a Chain.” *Bulletin de l’Academie Imperiale des Sciences de St. Petersburg* 6, no. 7 (1913).

Rugg, Gordon. “An Elegant Hoax? A possible solution to the Voynich manuscript.” *Cryptologia* 28, no. 1 (2004): 31–46.

Rugg, Gordon, and Gavin Taylor. “Hoaxing statistical features of the Voynich Manuscript.” *Cryptologia* 41, no. 3 (2017): 247–268.

Turchin, Peter, Jonathan M. Adams, Thomas D. Hall. “East-West Orientation of Historical Empires and Modern States.” *Journal of World-Systems Research* 12, no. 2 (December 2006): 219–229.

Verbunt, Frank, and Robert H. van Gent. “The star catalogues of Ptolemaios and Ulugh Beg, Machine-readable versions and comparison with the modern HIPPARCOS Catalogue.” *Astronomy & Astrophysics* (2012), 544, A31.

4. Monographs

Prinke, Rafał T. “History of the Manuscript from Rudolf to Voynich, presentation at the 100th anniversary of the discovery of the Voynich MS.” Monteporzio Catone, Italy: unpublished? May 11, 2012.

5. News articles

Friedman, Elizebeth. “The Most Mysterious MS. Still an Enigma.” *Washington Post*, August 5, 1962, E1.

ABBREVIATIONS

[…]	author's note or translation
[…]	keyboard assignment in v104 alphabet
[sic]	quoted exactly as in original
{…}	keyboard assignment in v101 alphabet
8vo	octavo
AD	Anno Domini (year of Christian era)
Alph.	alphabet
Alphab.	alphabet
AMS	Accelerator Mass Spectrometry
arch.	archaic
ASCII	American Standard Code for Information Exchange
BC	Before Christian (era)
BLAS	Basic Linear Algebra Subprograms
BoI	Bureau of Investigation
C	100 (Roman numeral)
c.	circa (approximately)
ca.	circa (approximately)
CC	200 (Roman numeral)
CCC	300 (Roman numeral)
CD	400 (Roman numeral)
cf.	comparable with
contd	continued
COUNTIF	conditional count function in Microsoft Excel
D	500 (Roman numeral)
DC	District of Columbia
DNA	deoxyribonucleic acid
DOCID	document identification number
EM	expactation maximization
EVA	Extended Voynich Alphabet
ex	excluding
f.	folio
f.	feminine
FBI	Federal Bureau of Investigation

FCSH	Faculdade de Ciências Sociais e Humanas
ff.	folios
FOIA	Freedom Of Information Act
G1	one-glyph strin
G2	two-glyph string
G3	three-glyph string
G4	four-glyph string
G5	five-glyph string
G6	six-glyph string
GBIF	Global Biodiversity Information Facility
HMM	Hidden Markov Model
I	1 (Roman numeral)
i.e.	id est (that is)
IAU	International Astronomical Union
II	2 (Roman numeral)
III	3 (Roman numeral)
IV	4 (Roman numeral)
IVTFF	Intermediate Voynich Transliteration File Format.
IX	9 (Roman numeral)
L	50 (Roman numeral)
L.	Linnaeus (classification)
LX	60 (Roman numeral)
LXI	61 (Roman numeral)
M	1000 (Roman numeral)
m.	masculine
M.S.	manuscript
MATTR	moving average type-token ratio
MD	Maryland
mm	millimeters
MMPND	Multilingual Multiscripted Plant Name Database
MS	manuscript
NASA	National Aeronautics and Space Admenistration
NCBI	National Center for Biotechnology Information
NOVA	Universidade NOVA de Lisboa
NSF	National Science Foundation
OCR	optical character recognition
OVI	Opera del Vocabolario Italiano
p.	page
pl.	plural
pp.	pages
RSQ	R squared (correlation coefficient) in Microsoft Excel

sing.	singular
sp.	species undetermined
spp.	species (plural) undetermined
St.	Saint
TTR	type-token ratio
US	United States of America
V	5 or 5th (Roman numeral)
v101	Voynich 101 alphabet (Claston)
v102	Voynich 102 alphabet (author)
v103	Voynich 103 alphabet (author)
v104	Voynich 104 alphabet (author)
VI	6 (Roman numeral)
VII	7 (Roman numeral)
VIII	8 (Roman numeral)
VM	Voynich manuscript
X	10 (Roman numeral)
XV	15 (Roman numeral)
XVI	16 (Roman numeral)
XVII	17 (Roman numeral)